A statistical investigation of fraud and misconduct in clinical trials

Sanaa Mohammed Al-Marzouki

London School of Hygiene &Tropical Medicine

Thesis submitted for degree of Doctor of Philosophy (Ph.D.)

ABSTRACT

Research misconduct can arise in any area of research and can discredit the findings. Research misconduct at any level is unacceptable, especially in a clinical trial. Because the results from clinical trials are used to decide whether or not treatments are effective, and affect decisions that may influence treatment choices for large numbers of patients, the prevention and detection of scientific misconduct in clinical trials is particularly important.

Chapter 1 outlines some definitions of research misconduct, discusses the underlying motivations behind it, and the overall prevalence of research misconduct beyond that occurring in clinical trials. Different ways to detect and prevent research misconduct are also presented. In addition, an initial insight into the types of scientific misconduct that have been reported as occurring in clinical trials, based on a search of the PubMed database between January 2000 and July 2003 is provided. Thirty-eight published reports were found, but they provide no indication of the relative importance of different types of scientific misconduct in clinical trials.

Chapter 2 presents a three-round Delphi survey aimed at achieving consensus among experts in clinical trials on what types of scientific misconduct are most likely to occur, and are most likely to influence the results of a clinical trial. This study identified thirteen forms of scientific misconduct for which there was consensus (>50%) that they would be likely or very likely to distort the results and consensus (>50%) that they would be likely or very likely to occur. Of these, the over-interpretation of 'significant' findings in small trials, selective reporting and inappropriate sub-group analyses were the main themes.

To prevent such types of misconduct in clinical trials, the issue of selective reporting of outcomes or sub-group analyses and the opportunistic use of the play of chance (inappropriate sub-group analyses) should be addressed. Full details of the primary and secondary outcomes and sub-group analyses need to be specified clearly in protocols. Any sub-group analyses reported without pre-specification in the protocol would need supporting evidence within the publication for them to be justified.

Chapter 3 explores selective reporting and inappropriate sub-group analyses within a cohort of randomised trial protocols approved by *the Lancet*. It determines the prevalence of selective reporting of primary and secondary outcomes and sub-group analyses in published reports of randomised trials. It also examines how sub-group analyses are described in protocols and how sub-group analyses are reported, and whether they match those specified in the protocol.

Of 56 accepted protocols, four non-randomized trials were excluded. For the remaining 52, permission to review them was obtained for 48 (92%). Of those 48 trials, 30 (63%) trials were published. This study identifies some shortcomings in the reporting of the results of primary and secondary outcomes and sub-group analyses. It shows at least one unreported primary, secondary or sub-group analysis in 37%, 87%, and 50% of the trials, respectively. It also shows that the pre-specification and reporting of sub-group analyses are often incomplete and inaccurate. The majority of protocols gave hardly any detail on this matter. There was notable deviation from the protocols in reports in several of the trials.

Data fabrication and falsification were judged by the experts in the Delphi survey to be unlikely to occur. However, they can have major effects on the outcomes of clinical trials if it they do occur. A systematic review was conducted in chapter 4, to identify the available statistical techniques that could be used for the detection of data fabrication and falsification.

Chapter 5 examines the ability of these statistical techniques to detect data fabrication in data from two randomised controlled trials. In one trial, the possibility of fabricated data had been raised by British Medical Journal (*BMJ*) referees and the data were considered likely to contain fraudulent elements. For comparison, a second trial, about which there were no such concerns, was analysed using the same techniques, and no hint appeared of any unusual or unexpected features was shown.

Finally, chapter 6 contains some concluding remarks, a discussion of the strengths and weaknesses of this research and suggestions for future research.

Table of contents

List of Tables

List of Figures

Acknowledgement

My ardent gratitude to God the benevolent,

Who has given me loving parents,

Supportive friends,

&

Opportunities to learn

I wish to express my sincere thanks to both of my supervisors, Professor Ian Roberts and Mr Tom Marshall (London School of hygiene & Tropical Medicine) for their invaluable support and advice over the last three years.

I would also like to thank Professor Stephen Evans for his valuable suggestions and for reviewing the manuscript of the thesis.

I am very grateful for the love and support of my parents, my sister and my brothers.

The financial support of King Abdulaziz University in Kingdom of Saudi Arabia is gratefully acknowledged.

CHAPTER 1

Background

Misconduct is a serious problem in research and any form of misconduct can discredit the findings of that research. It jeopardises scientific reliability and erodes the trust and confidence of the public. Research misconduct may, and does, occur in many disciplines, such as physics (1), nano-electronics (2), ecology (3) as well as in clinical trials (4).

Organisations conducting research need an internal or external framework of good practice, guidance, policies, research monitoring and auditing. Policies and guidelines for research monitoring and auditing will help to deter research misconduct, and importantly, help identify inadequate research practices before they become cases of research misconduct. A research culture of good conduct will also help reduce levels of misconduct. Organisations should have sufficient procedures to identify misconduct and should have clear procedures for handling cases of alleged or suspected research misconduct.

1.1 Definitions of research misconduct

Definitions of research misconduct are needed as a basis for assessment of how commonly scientific research misconduct occurs. Unfortunately, it is not easy to arrive at a comprehensive and precise definition. Smith (5) suggests that an operational definition is almost unachievable. Various definitions of research misconduct have been produced by various organisations.

In 1995, the United States Commission on Research Integrity defined it as "significant misbehaviour that improperly appropriates the intellectual property or contributions of

others, that intentionally impedes the progress of research, or that risks corrupting the scientific record or compromising the integrity of scientific practice. Such behaviours are unethical and unacceptable in proposing, conducting, or reporting research or in reviewing the proposals or research reports of others" (6).

The US federal government produced a slightly shorted definition in 2000 (7): "Research misconduct is defined as fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results".

Fabrication is making up data or results, and then recording or reporting the made up results. Falsification is manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record. Plagiarism is the appropriation of another person's ideas, processes, results or words without giving appropriate credit. It is clear that research misconduct does not include honest error or differences of opinion.

The Nordic countries and Britain decided on broad definitions (8). In 1992, the Danish Committee on Scientific Dishonesty used the terms "Intention or gross negligence leading to falsification of scientific message or a false credit or emphasis given to a scientist" (8).

The Norwegian Committee on Scientific Dishonesty proposed the following definition in 1994: "All serious deviation from accepted ethical research practice in proposing, performing, and reporting research"(8).

In 1998, the Finnish Committee of Scientific Dishonesty defined research misconduct as "Presentation to the scientific community of fabricated, falsified, or misappropriated observations or results and violation against good scientific practice" (8).

Also in 1998, the Swedish Committee of Scientific Dishonesty proposed the definition: "Intentional distortion of the research process by fabrication of data; theft or plagiarism of data, text, hypothesis, or methods from another researcher's manuscript or application from or publication; or distortion of the research process in other ways" (8).

The definitions of the Nordic countries are broad and include a range of practices. Intention to deceive is the link in all four countries.

A British consensus statement developed at a UK Consensus Conference on Misconduct in Biomedical Research organised by the Royal College of Physicians of Edinburgh in 2000 defined research misconduct as "Behaviour by a researcher, intentional or not, that falls short of good ethical and scientific standards" (9). This definition does not depend on intention and does not include qualification about falling '*seriously*' short of good standards.

1.2 Motivations behind research misconduct

It is useful to understand the motivation to commit misconduct, because knowledge about motivation could contribute to a solution to the problem. However, the motivations to commit misconduct are as varied as human personalities. According to Taylor *et al* (10), there are three motivating factors in research misconduct. These motivators are: laziness, financial gain, and desire for professional recognition. Similar factors that motivate scientists to commit misconduct were discussed by Brock (11) and briefly are:

- Pressure to publish: scientists need to publish new articles continuously. This strong imperative to publish may motivate desperate scientists to commit some kind of research misconduct.

- Unreasonable expectations: a delay on registering and approval of a new product can lead to loss of sales profits of many millions of pounds. In the light of this pressure, it is not very surprising to find cases of shortcuts in the development process. These shortcuts may lead to fraud.
- Greed: many kinds of misconduct have been linked to studies involving sums of money. For example, when investigators are given payments to recruit or follow up patients, this may be a cause of fraud.

These may be the most common motivations, but there are some other not so commonly factors cited by Duff (12). Every case is different thus, the cures for every case may also be different.

1.3 Prevalence of research misconduct

It is difficult to quantify the extent of research misconduct, because many cases remain discovered or unreported, even if there are central databases for reported cases of research misconduct. Exact estimates of the prevalence of misconduct vary considerably (13-17). Institutions in Britain have roughly one serious case a year, which means about 50 a year nationally (18).

Several surveys have shown that a majority of researchers know of cases of misconduct, and that many of these cases have never been reported (17). The proportion of investigators who had actually committed fraud was less than 1% (13).

A questionnaire study to determine the prevalence of observed and personal research misconduct among newly appointed medical consultants in the Mersey region, United Kingdom, was conducted between Jan 1995 and Jan 2000 in seven different hospital trusts

(19). From 194 respondents (a response rate of 63.6%), 56% had observed some form of research misconduct; 5.7% of respondents admitted to past personal misconduct; 18% of respondents were either willing to commit, or unsure of their willingness to commit possible research misconduct in the future.

This survey showed awareness of a higher prevalence of observed misconduct (56%) among newly appointed consultants in the UK than in a comparable study (36%) in California (20). In the American survey on the prevalence of research fraud directed at biomedical trainees, 15% of the respondents admitted personal instances of misconduct. Such differences may vary according to the different study populations.

Another recent survey in the US (21) covered several thousands of scientists funded by the National Institutes of Health (NIH), anonymously, and reported the replies from the 3247 who responded (52% response rate). Just 0.3% of those scientists confessed that they had falsified or "cooked" research data, 1.4% admitted to plagiarism and almost a seventh (15.3%) indicated that they had dropped observations or data points from analysis based on a gut feeling. Lesser violations were for more common misconduct, including 4.7% who admitted to publishing the same data in two or more publications to beef up their résumé's, 13.5% who had used inadequate or inappropriate research designs and 15.5% who admitted that they had changed the design, methodology or results of a study in response to pressure from a funding source. These findings are based on self reporting of behaviour, which is likely to mean under-reporting and conservative estimates, despite assurances of anonymity.

The incidence of misconduct in multi-centre studies revealed by audits was low, with figures of 0.29% for the US (22), 0.4% for the UK (23), and 0.43% for Europe and South Africa (14). The prevalence of misconduct is difficult to deduce from surveys with any certainty,

but their findings add an impression to the public that research misconduct is a frequent occurrence. In essence, research misconduct may be much more common than reported, because of concealed cases.

1.4 Dealing with research misconduct

The United States and some European countries have set up national bodies that have strategies and policies to deal with the problem of scientific misconduct. Members of these bodies are scientifically and legally qualified. Other countries, including the UK, still have no coherent system, and lack processes to deal with scientific misconduct.

The Office of Research Integrity (ORI) is one of the bodies concerned with research integrity in the United States (24). The ORI promotes research integrity, writes policies and regulations to prevent and detect scientific misconduct, gives advice, supervises institutional investigations of research misconduct and facilitates the responsible conduct of research through educational, preventive and regulatory policies. It regularly publishes its findings.

In the UK, research misconduct has been discussed and guidelines on how to deal with cases of misconduct were suggested by the Royal College of Physicians of London in 1991 (25). Guidance on research governance, which is applicable to the NHS and universities was also published by the Department of Health in 1997 (26).

The Committee on Publication Ethics (COPE) was established in 1997 to help editors of medical journals to respond to concerns about the integrity of studies submitted to them. It was founded by British medical editors from the BMJ, GUT and *the Lancet*. The first aim of COPE was to advise on cases brought by editors. Other aims were to publish an annual report describing those cases of misconduct, and produce guidance on good practice. So far,

COPE has described around 250 cases, all of which are explained anonymously in the committee's annual reports (18).

In 2006, the UK panel for health and biomedical research integrity initiated a new independent body to tackle biomedical research misconduct. Promoting good practice and providing advice to universities, the NHS and industry are the main tasks of the body, not to investigate misconduct itself (27). Supporting whistleblowers who report or allege cases of fraud and offering training, seminars and advice on how to investigate cases of misconduct are other duties of this body.

The National Research Ethics Council of Finland was founded in 1991 (8). The council is subordinate to the education ministry. In 1998, the Council produced guidelines for the prevention, handling and investigation of misconduct and fraud in scientific research, which allow universities and research institutes to prevent and to investigate alleged cases of misconduct. The Council is informed of all inquiries and receives a final report on each case from the investigating institution. The Council does not produce legally binding decisions, but it has an advisory role. In 1999, the Council received 19 cases of suspicion of misconduct and 10 cases in 2000 (28).

The Danish Committee on Scientific Dishonesty was established in 1992 (8). It is able to investigate cases, express its opinion, and deal with aspects of scientific dishonesty in medical science. This committee continued its work until the end of 1998. From 1992 to 1998, the committee received 45 claims of alleged misconduct and investigated 25 cases, but only four of them were confirmed to include misconduct. A new committee system was instituted in 1999 to cover fraud in all scientific fields and to handle cases concerning scientific misconduct. The various committees publish annual reports of cases of misconduct

anonymously. In 1999 and 2000, allegations were submitted to the committee related to the field of social science and the humanities.

The National Committee for the Evaluation of Dishonesty in Health Research in Norway was established in 1994 (8). The Committee consisted of representatives from several healthcare professions and a judge. The tasks assigned to the Committee are to prevent scientific dishonesty, to ensure the investigation of alleged incidents of misconduct reported in the health sciences, and to inspect. Between 1994 and 2000, 11 cases of suspected misconduct were accepted for investigation (29).

A national committee for research ethics was set up in Sweden in 1997 (8). Recommendations were put forward in 1999 in a report entitled "Good Practice in Research". In order to increase public overseeing of research systems. It included a National Commission to deal with allegations of research misconduct. The investigations in Sweden were along the lines of the Danish approach.

The scandal of the two cancer researchers, Friedhelm Herrmann and Marion Brach, who fabricated data in about 47 papers in Germany in 1997, led the German Research Foundation (30) to:

- Appoint an ombudsman for science, in 1999, who could advise and assist scientists in questions of good scientific practice;
- Appoint an international commission with the mandate to explore the causes of misconduct, to take preventive measures, and to make recommendations on how to safeguard future research.

France has been absent from the debate concerning official sanction of individual cases of fraud, perhaps because of a desire to hide such problems or because of the absence of well-codified rules. Nevertheless, the French national institute, INSERM, which is responsible for research in the biomedical and health fields, does investigate allegations of scientific misconduct. Recommendations on scientific integrity were made in 1998 by INSERM. In 1999, 18 cases of alleged research misconduct were investigated and 25 cases in 2000 (31).

These institutional arrangements are undoubtedly important in ensuring good conduct and in detecting misconduct. These committees cannot take definitive decisions on whether dishonesty has taken place or not. A body to investigate allegations, a fair system for reaching judgments and an adequate training system for teaching good practice are needed.

Not all countries have reported setting up such arrangements; the lack of reports from Asian countries is to be noted. However, how effective the approach is, overall, is not clear, and it is hard to see how to proceed further in general terms.

It is important not to ignore the role of the "whistleblower", and a system where whistleblowers are not penalised for a truthful declaration is important. The various committees referred to above can facilitate whistleblowing, as can ready access to a free press.

In clinical trials, monitoring and arrangement committees are also able to act, as discussed in section 1.6.1.

1.5 Main objectives and approaches

This thesis aims to identify statistical techniques that may be employed to detect fraud and scientific misconduct in clinical trials. The following steps are to be carried out:

1 Listing and grading the principal types of scientific misconduct which may arise in clinical trials. To set this issue in context, two approaches are followed; (i) reviewing the scientific literature over a short three-year period; (ii) conducting a survey of experts' opinion regarding this issue.

2 Examining further the important types of scientific misconduct that are identified by the experts and asking them to assess their potential impact on the results of the clinical trials.

3 Identifying the statistical techniques that may be used to detect data fabrication and falsification. Producing recommendations as to when and how these techniques should be used.

4 Checking the use of some of the techniques on two real datasets to demonstrate fabrication or its absence.

1.6 Clinical trials

A clinical trial is a planned experiment designed to assess the efficacy of a treatment in humans by comparing the outcomes in a group of patients treated with the test treatment with those observed in a comparable group of patients receiving a control treatment. Patients in both groups are enrolled, treated and then followed over the same period. In a randomised controlled trial, participants are allocated to the groups at random, and in a single blind or double blind trial, precautions are taken to ensure that the participants and/or the persons involved in their treatment and in handling the data do not know to which group the individual participants belong.

Although any form of misconduct can discredit the findings of a clinical trial, misconduct that distorts the estimate of the treatment effect or the assessment of statistical significance

is of special importance, since it may lead to patients being given useless or harmful treatments or to patients being denied effective treatments. Scientific misconduct at any level is unacceptable, since it jeopardises scientific integrity, endangers patients and erodes the trust and confidence of the public.

1.6.1 Scientific misconduct in clinical trials

As in any other area of research, scientific misconduct can arise in clinical trials. However, because the results from clinical trials are used to decide whether or not treatments are effective (decisions that may influence treatment choices for large numbers of patients) the detection and prevention of scientific misconduct in clinical trials have great importance. In clinical trials, Trial Steering Committee and Trial Management Groups should be set up for each project as appropriate, and these should have terms of reference that include research misconduct.

Monitoring visits to the clinical centres participating in a trial is one approach to fraud control (32, 33). In some circumstances, such monitoring should be routine, and some instances of fraud have been detected during these visits (34). However, it is expensive and difficult to verify everything, especially when the volume of data to be checked is very large. Where there are particular grounds to suspect misconduct, it is of help to submit clinical trial data to more extensive checks.

Statistical techniques for fraud detection can be used as screening mechanisms or for further investigation of data that fall under suspicion. They can be implemented more easily than the monitoring approach, especially with modern computer programs for statistical data analysis, so long as the primary data can be obtained. An excellent way of checking fabricated data is on the basis that humans that are unable to generate long sequences of

numbers that pass simple tests for randomness (35). Terminal digit preference may easily reveal data fabrication (36). In clinical trials, there are two measures that may be particularly effective in preventing misconduct. These two measures are:

- A simplification of the eligibility criteria, because some misconduct may occur if the eligibility criteria are excessively restrictive (37-39).
- Allowance for missing data. Although complete data are certainly better than missing data, missing data should generally be accepted in clinical trials (though not for the primary endpoint of the trial). Attempting to collect too much data and repeatedly demanding complete data on all patients may lead to fraud, rather than prevent it.

1.6.2 Various types of scientific misconduct reported in clinical trials

To provide an initial insight into the types of scientific misconduct that may occur in clinical trials and are reported in the scientific literature, a limited search using the PubMed database between January 2000 and July 2003 was carried out, using the key words "clinical trials" AND "scientific misconduct". The search was limited to human subjects and the English language.

Fifty-seven papers were found in the search, thirty-eight of which (67%) reported an investigation of alleged scientific misconduct in twenty-four clinical trials. Another sixteen (28%) papers discussed general topics related to scientific misconduct, such as definition, types of scientific misconduct or advice and guidelines for researchers and some ethical issues. The text of the remaining three papers (5%) was not located through an inter-library search.

The pie chart (Figure 1.1) shows the stage of the trial in which the scientific misconduct was thought to have arisen. (The denominator here is the number of instances of misconduct reported). The 38 papers (40-77) cited 23 trials with fifty-two instances of scientific misconduct reported, of which 23% were in the design, 46% in the conduct, 4% in the analysis and 27% in the reporting stage.

Figure 1.1 Stage of clinical trial in which scientific misconduct occurred

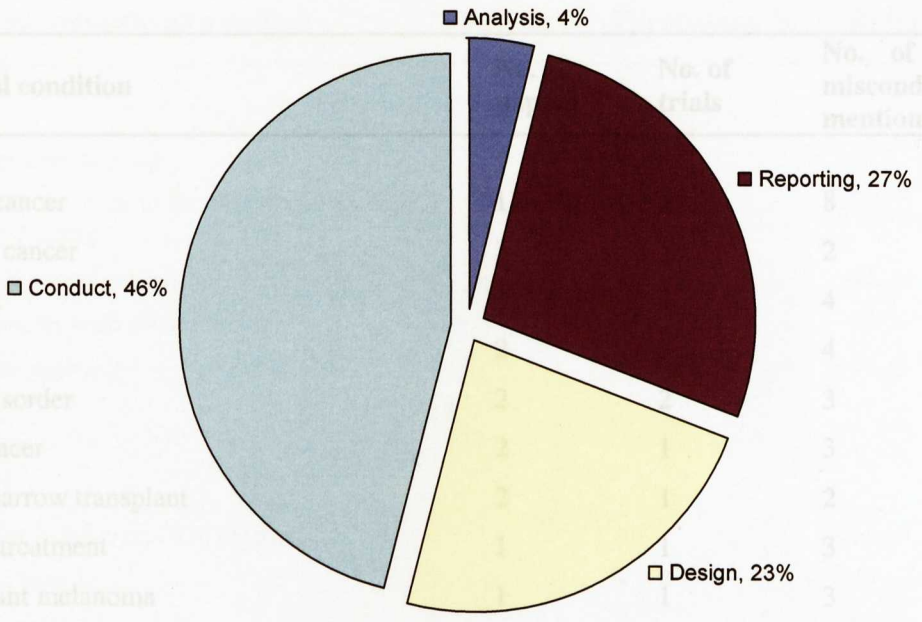


Table 1.1 Reporting of scientific misconduct in clinical trials by disease area

Medical condition	No. of papers	No. of trials	No. of cases of misconduct mentioned†
Breast cancer	11	1	8
Kidney cancer	3	1	2
Asthma	2	1	4
HIV	2	2	4
Gene disorder	2	2	3
Oral cancer	2	1	3
Bone marrow transplant	2	1	2
Cancer treatment	1	1	3
Malignant melanoma	1	1	3
Cardiovascular secondary prevention	1	1	2
Neonatal continuous negative extra-thoracic pressure	1	1	1
Neurotoxin effect of the excitatory amino acid glutamate in the eye	1	1	1
Meningitis	1	1	1
Lead paint	1	1	1
Stroke	1	1	1
Coronary artery disease	1	1	1
Unspecified	5	5	12
Total	38	23	52

†Each case may be reported more than once for the same trial, but one case is counted.

Table 1.2 The percentage of different forms of scientific misconduct mentioned in the reviewed papers, classified by type

Types of scientific misconduct	No. of cases	Percentage %	Reference
Design			
Unethical control group	1	8	53
Not telling the truth to the patients about the source of funding	1	8	50, 51
No test on animal models	2	17	42, 52
Not telling the truth about the risk	2	17	49-51
No ethical approval	6	50	40-48
Total	**12**		
Conduct			
Sloppiness in the practice of the trial	1	4	65,66
Fictitious patients	1	4	67
Missing medical record	2	8	62,64
Deviation from the protocol regarding eligibility	3	13	52,54,61-63
Fabrication of data	3	13	58-60
Falsification of data	6	25	54-56, 58,59,68
No informed consent	8	33	40,42,43,46,53-57
Total	**24**		
Analysis			
Deviation from the analysis plan	1	50	70
Ghost analysis	1	50	69
Total	**2**		
Reporting			
No peer review	1	7	73
Duplicate publication	1	7	72
Copying results	1	7	69
Misrepresenting results	1	7	45,54,59,74,75
Publishing early positive results	1	7	76,77
Fabricating and manipulating results	2	14.5	55,58
Ghost writing	2	14.5	69,72
Incomplete report	2	14.5	63,70
Not reporting adverse events	3	22	40,52,71
Total	**14**		

Table 1.1 summarises the information extracted from the papers. The first column lists the medical condition for which there were reports of scientific misconduct in clinical trials. The second column indicates the numbers of papers reporting scientific misconduct for each medical condition. For example, in the asthma trials, two papers were found for one trial. From these, four types of scientific misconduct were mentioned.

The same trial is sometimes discussed in several papers. In the area of kidney cancer, three papers reported on the same trial (65, 66, 68). These papers report an investigation by a university in Gottingen into alleged scientific misconduct in the study, which was led by Alexander Kugler, and which claimed that kidney cancer could be treated using a vaccine made from a tumour cell fused with a healthy dendritic cell from the immune system. The university said that its investigator had found evidence of sloppiness that constituted misconduct; the data in the study were handled incorrectly, and there was fabricated data.

In the areas of breast cancer, the scientist, Werner Bezwoda claimed in a study, conducted in South Africa, that high-dose chemotherapy prolonged the lives of some women with advanced breast cancer. This study was published in the Journal of Clinical Oncology during 1995, but the journal retracted the article. There were eleven papers discussing the Bezwoda study, which reported eight different types of scientific misconduct (44-47, 54, 59 61, 62, 74-76).

Scientific misconduct was considered to occur most commonly (46%) during the conduct of the trial (see Table 1.2). Of these instances of misconduct, 38% involved data falsification and fabrication. A further 33% and 13% of the 24 reported episodes of scientific misconduct respectively were due to failure to obtain informed consent and changes in the inclusion criteria.

Twenty-seven percent of the reported instances of scientific misconduct occurred in the reporting process. Of these, 22% involved failure to report adverse events. The remaining 78% are divided into eight different forms of misconduct. Twenty-three percent of the reported scientific misconduct was during the design stage. Of these, fifty percent of episodes involved a lack of ethics committee approval. A minority of reported episodes (4%) of misconduct were considered in the analysis stage.

The small group of papers referenced above has identified several different types of scientific misconduct and how often they occur in clinical trials. However, it provides insufficient coverage and detail of the overall situation or the magnitude of effect of each type of misconduct reported. Another approach to eliciting expert opinions is to use a Delphi survey, as presented in the next chapter, to assess which types of scientific misconduct are most likely to distort the results of a clinical trial.

References

1 Giles J. Plagiarism in Cambridge physics lab prompts calls for guidelines. *Nature* 2004; 427:3.

2 Brumfiel G. Time to write up? *Nature* 2002; 418:120–1.

3 Abbott A. Prolific ecologist vows to fight Danish misconduct verdict. *Nature* 2004; 427:381.

4 Al-Marzouki S, Evans S, Marshall T, Roberts I. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *BMJ* 2005; 331:267-270.

5 Smith R. What is research misconduct? *J Roy Coll Physicians Edin.* 2000; 30:4 - 8.

6 Integrity and Misconduct in Research. Report of the Commission on Research Integrity to the Secretary of Health and Human Services, the House Committee on Commerce, and the Senate Committee on Labor and Human resources. 1995. [http://gopher.faseb.org/opar/cri.html].

7 Office of Science and Technology Policy, Executive office of the President. *Federal Policy on Research Misconduct. Federal Register* 2000; 76260-4[http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=2000_register&docid=00-30852-filed] Accessed July 2003.

8 Nylenna M, Andersen D, Dahiquist G, Sarvas M, Aakvaag A. Handling of scientific dishonesty in the Nordic countries. *Lancet* 1999; 354:57 -61.

9 Nimmo WS. Joint Consensus Conference on Misconduct in Biomedical Research. *Proceedings of the Royal College of Physicians of Edinburgh* 2000; 30 (Supplement 7).

10 Taylor R, McEntegart D, Stilman E. Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. *Drug Information Journal* 2002; 36:115-125.

11 Brock P. A pharmaceutical company's approach to the threat of research fraud. In: Lock S, Wells F, Farthing M, eds. *Fraud And Misconduct In Biomedical Research*, 3rd ed. London: BMJ Publishing Group; 2001:89-104.

12 Duff G. The researcher perspective. *Proceedings of the Royal College of Physicians of Edinburgh* 2000; 30:26.

13 Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B, *et al.* The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Statistics in Medicine* 1999; 18: 3435-51.

14 Schmidt J, Gertzen H, Aschenbrenner KM, Ryholt-Jensen S. Detection fraud using auditing and biometrical methods. *Appl Clin Trials*. 1995; 4:40-49.

15 O'Donnell P. Facing up to fraud. *Appl Clin Trials*. 1993; 2:36-40.

16 Well F. Investigating fraud-again. *Appl Clin Trials*. 2000; 10:26-27.

17 Ranstam J, Buyse M, George SL, Evans S, Geller NL, Scherrer B, et al. Fraud in medical research: an international survey of biostatisticians. ISCB Subcommittee on Fraud. *Controlled clinical trials* 2000; 21:415-27.

18 Smith R. Research misconduct: the poisoning of the well. *J R Soc Med* 2006; 99:232-237.

19 Geggie D. A survey of newly appointed consultant's attitudes towards research fraud *Journal of medical ethics* 2001; 27:344-6.

20 Kalichman M, Friedman P. A pilot study of biomedical trainees' perceptions concerning research ethics. *Academic Medicine* 1992; 67:769–75.

21 Martinson BC, Anderson MS, deVries R. Scientists behaving badly. *Nature* 2005; 435: 737-8.

22 Weiss RB, Vogelzang NJ, Peterson BA, Panasci LC, Carpenter JC, Gavigan M, *et al.* A successful system of scientific data audits for clinical trials. *JAMA* 1995; 270: 459-464.

23 Hone J. Combating fraud and misconduct in medical research. *Scrip Magazine* 1993; 14-15.

24 Rennie D, Gunsalus CK. Regulations on scientific misconduct: lessons from the US experience. In: Lock S, Wells F, Farthing M, eds. *Fraud And Misconduct In Biomedical Research*, 3rd edn. London: BMJ Publishing; 2001:13-31.

25 Working Party. Fraud and misconduct in medical research Causes, investigation and prevention. London: Royal College of Physicians; 1991.

26 Medical Research Council. Policy and procedure for inquiring into allegations of scientific misconduct. MRC Ethics Series. London: MRC; 1997.

27 Cole A. UK launches panel to tackle research misconduct. *BMJ* 2006;332:871.

28 Launis V. Scientific fraud and misconduct in Finland. In: Lock S, Wells F, Farthing M, eds. *Fraud and Misconduct in Biomedical Research*, 3rd ed. London: BMJ Publishing Group; 2001:131-133.

29 Nylenna M. experiences of fraud and misconduct in healthcare research in Norway. In: Lock S, Wells F, Farthing M, eds. *Fraud and Misconduct in Biomedical Research*, 3rd ed. London: BMJ Publishing Group; 2001:134-139.

30 Stegemann-Boehl S. Dealing with misconduct in science: German efforts. In: Lock S, Wells F, Farthing M, eds. *Fraud and misconduct in medical research.* 3rd ed. London: *BMJ* Publishing Group, 2001:140-151.

31 Husson J, Demarez J. Fraud and misconduct in medical research in France. In: Lock S, Wells F, Farthing M, eds. *Fraud And Misconduct In Biomedical Research*, 3rd ed. London: BMJ Publishing Group; 2001:152-169.

32 Mackintosh DR, and Zepp VJ. Detection of negligence, fraud, and other bad faith efforts during field auditing of clinical trial sites. *Drug Information Journal* 1996; 30:645-653.

33 Schwarz RP. Maintaining integrity and credibility in industry-sponsored clinical research. *Controlled Clinical trials* 1991;12:753-760.

34 Seachrist L. NIH trial monitoring: hit or miss? *Science* 1994; 264: 1534-1537.

35 Rao CR. Statistics and truth, International Co-operative Publishing House, Burtonsville, 1989.

36 Preece DA. Distribution of final digits in data. *Statistician* 1981; 30: 31- 60.

37 Dingell JD. Shattuck lecture: Misconduct in medical research. *New England Journal of Medicine* 1993; 328:1610-1615.

38 Yusuf S, Held P, Teo K, Toretsky ER. Selection of patients for randomized controlled trials: implications of wide or narrow eligibility criteria. *Statistics in Medicine* 1990; 9:73-86.

39 George SL. Reducing patient eligibility criteria in cancer clinical trials. *Journal of Clinical Oncology* 1996; 14:1364-1370.

40 Report on research death says review board is overworked. *The New York times* 2001; 31:A16.

41 Johns Hopkins is investigating study in India by a professor. *The New York times* 2001; A15.

42 US university finds scientist flouted rules for clinical trials. *Lancet* 2001; 358:1791.

43 Sharma DC. Research halted at Indian centre accused of misconduct. *Lancet* 2001; 358:992.

44 Jayaraman KS. Johns Hopkins embroiled in fresh misconduct allegations. *Nature* 2001; 412:466.

45 Scientific misconduct. Cancer researcher sacked for alleged fraud. *Science* 2000; 287:1901-2.

46 High-dose chemotherapy for high-risk primary breast cancer: an on-site review of the Bezwoda study. *Lancet* 2000; 355:999-1003.

47 Cleaton JP. Scientific misconduct in a breast-cancer chemotherapy trial: response of University of the Witwatersrand. *Lancet* 2000; 355:1011-2.

48 Sharma DC. Indian medical agencies probe illegal VEGF trials. *Lancet* 2001; 357:1422.

49 Lewin T. U.S. investigating Johns Hopkins study of lead paint hazard. *The New York times* (Print). 2001; A11.

50 McCarthy M. US newspaper charges research centre with misconduct. *Lancet* 2001; 357:865.

51 Marshall E. Clinical research Fred Hutchinson Center under fire. *Science* 2001; 292:25.

52 Lemonick MD, Goldstein A. At your own risk some patients join clinical trials out of desperation, others to help medicine advance. Who is to blame if they get sick--or even die? *Time*. 2002; 159:46-56.

53 Todres J. Can research subjects of clinical trials in developing countries sue physician-investigators for human rights violations? *New York Law School journal of human rights* 2000; 16:737-68.

54 Cauvin HE. Cancer researcher in South Africa who falsified data is fired. *The New York times* 2000:A16.

55 Koenig R. Scientific misconduct Fallout from German fraud case continues. *Science* 2001; 291:1876-7.

56 Hoeksema HL, Troost J, Grobbee DE, Wiersinga WM, Wijmen FC, Klasen EC. Fraud in a pharmaceutical trial. *Lancet* 2000; 356:1773.

57 Hey E, Chalmers I. Investigating allegations of research misconduct: the vital need for due process. *BMJ* 2000; 321:752-5.

58 Beach J. Clinical trials integrity: a CRO perspective. *Accountability in research* 2001; 8:245-60.

59 Gralow JR, Livingston RB. University of Washington high-dose cyclophosphamide, mitoxantrone, and etoposide experience in metastatic breast cancer: unexpected cardiac toxicity. *Journal of clinical oncology* 2001; 19:3903-4.

60 Dalton R. Private investigations. *Nature* 2001; 411:129-30.

61 Farham B, Bradbury J. Suspicions raised over breast-cancer-therapy trial. *Lancet* 2000; 355:553.

62 Gottlieb S. Breast cancer researcher accused of serious scientific misconduct. *BMJ* 2000; 320:398.

63 Hilts PJ. FDA says researcher failed to report a second death linked to gene therapy. *The New York times* 2000; A20.

64 Ahmad K. Nigerian government investigates Pfizer drug trial allegations. *Lancet* 2001; 357:129.

65 Kerstholt M. No end in sight for German misconduct probe. *Nature* 2002; 417(6884):6.

66 Bostanci A, Vogel G. Research misconduct: German inquiry finds flaws, not fraud. *Science* 2002; 298:1531-3.

67 Cohen SN. Dipyridamole plus aspirin in cerebrovascular disease. *Archives of neurology* 2000; 57:1086-7.

68 Schiermeier Q. Cancer researcher found guilty of negligence. *Nature* 2002; 420:258.

69 Kaushansky K. Removing the cloud from industry-sponsored, multi-centred clinical trials. *Blood* 2001; 98:2001.

70 Lurie P, Wolfe SM. Outcomes of a trial of HIV-1 immunogen in patients with HIV infection. *JAMA* 2001; 285:2193-5.

71 Gelsinger P. Jesse's intent. *Bulletin of Medical Ethics* 2002; 179:13-20.

72 Ethical behaviour of authors in biomedical journalism. *Ann R Coll Physicians Surg Can* 2002; 35:81-5.

73 Cohn F, Manetta A. A teachable moment: research ethics revisited. *American journal of obstetrics and gynecology* 2003; 188:2.

74 Abratt RP, Vorobiof DA, Goedhals L, Rappaport B, Ruff P, Slabber CF. Scientific fraud and international research. *Journal of Clinical Oncology* 2001; 19:3592.

75 Alleged misconduct in breast cancer trial. *European J. of Cancer* 2000; 36:552.

76 Vorobiof DA, Abratt R, Rapoport B, Ruff P, Slabber C, Goedhals L. Scientific misconduct in cancer clinical trials. *South African Medical Journal* 2001; 91:614-5.

77 Antman K. Randomized trials of high-dose chemotherapy in breast cancer: fraud, the press and the data (or lessons learned in medical policy governing clinical research). *Transactions of the American Clinical and Climatological Association* 2002; 113: 56-66; discussion 66 -7.

CHAPTER 2

Scientific misconduct in clinical trials: a Delphi survey

2.1 Introduction

In the previous chapter, although different types of scientific misconduct were identified in clinical trials, there is currently little information about which types of scientific misconduct are most likely to distort the results of, or conclusions from, clinical trials and which are most likely to occur. With this important information, it would then be possible to provide guidelines to control, detect, and prevent these types of misconduct. To achieve this, I decided that a synthesis of expert opinion would provide highly relevant information, so the Delphi technique (1) was used. I chose this technique because people in groups tend to share a particular knowledge base, and they often provide a rich and valued resource for gaining further insights. The decisions from this technique are based on opinion from experts and the result is considered to be more reliable than individual statements and to be more objective in its outcomes (2,3).

2.2 The Delphi technique

The Delphi technique is a method for structuring a group communication process that is effective in allowing a group of individuals, as a whole, to deal with a complex problem. Most authors draw on all or some (4-6) of the definitions set out by Linstone and Turoff (1). The technique is an appropriate one to obtain the opinion of experts without necessarily bringing them together face to face.

The essence of the technique is straightforward. It comprises a series of questionnaires sent either by mail or via computerized systems, to a pre-selected group of experts. The

questionnaires are designed to elicit and develop individual responses to the problems posed and to enable the experts to refine their views. The main intent behind the Delphi technique is to overcome the disadvantages of conventional committee action by removing opportunities for individual domination and other problems in inter-personal interaction. It also overcomes the practical difficulties of arranging group meetings of busy people.

2.2.1 Characteristics of the Delphi technique

According to Fowles (7) the Delphi technique is characterized by anonymity, controlled feedback, and statistical summarisation.

Anonymity: Provides an equal chance for participants to present and express their opinions without feeling pressured psychologically by more influential panel members (8).

Controlled feedback: Through several rounds of the procedure; a summary of responses to the previous round is fed back to the panellists, which allows participants to modify their opinions regarding the consensus generated in previous rounds.

Statistical summarisation: Responses are classified and summarized statistically for each member and reflected as the final response. Often opinions falling in the bottom or top extremes are asked to give the group extra justification.

2.3 Methods

A Delphi survey is conducted in three stages: selection of the panel, rounds of enquiry, and analysis.

2.3.1 The expert panel

The first stage is the formation of a panel of experts. An expert may be considered as one of a group of informed individuals (9), as a specialist in the field (10) or someone who has knowledge about a specific subject (11-13). The use of experts has been criticised (14) as there is a potential for bias in the selection since the exact composition of the panel can affect the results obtained. However, another opinion stated that by having as diverse panels as possible, biases are able to be minimized (15,16).

There is wide variation in the numbers of participants in the surveys that have been carried out. Carley et al (17) notes that, panel sizes varying between less than 22, to more than 2000 (18). In this study, 40 experts in clinical trials were assembled from the list of people invited to respond to the UK Medical Research Council (MRC) Clinical Trials for Tomorrow consultation (19). These panel members were selected on the basis of their knowledge of the subject area and their willingness to be involved in research as is recommended when using the Delphi approach (20).

Each expert was sent a letter explaining the aims and methods of the study and asked if they would be willing to be considered for the expert panel and to take part in a Delphi survey with three rounds.

2.3.2 The Delphi Rounds

Data for the survey were collected using a three round Delphi process. The literature reports that participants often become fatigued after three rounds (21).

Round 1

The first round is completely unstructured and seeks an open response, thus allowing participants complete freedom in their responses on the topic under investigation (8).

Each participating expert was asked to list, briefly and concisely, four suggestions about how scientific misconduct can arise in the design, conduct, analysis and reporting of a clinical trial.

The returned questionnaires revealed a wide range of suggestions about scientific misconduct. These suggestions were collated and duplicates were removed from the list in preparation for the second round.

Round 2

Based on the responses from *round 1*, the list of collated suggestions was sent to each participant, whether or not they had responded to the first round. Participants were asked to rate each type of scientific misconduct on two dimensions: (1) the likelihood that it would occur in a clinical trial and (2) the likelihood that it would distort the results (i.e. have an effect on the magnitude of the treatment effect). Participants rated each suggestion on a five point scale from 1, "very unlikely" to 5, "very likely". Again, the responses were collated and summarised and the results were fed back to each participant.

Round 3

For round three, a list was prepared of all the types of misconduct from *round 2*, showing the frequency distributions of the scores for both dimensions. Each participant's response in the second round was indicated in that person's list under the appropriate number on the frequency distribution. Each participant was offered the opportunity to change his or her response in the light of the group's opinion by ticking a new value for the score or if they did not wish to change their opinion to tick the same number as before.

2.3.3 Analysis

There is no standard threshold for consensus in Delphi studies, and indeed this is often cited as a major deficiency in studies using this technique (22). Boyce et al (23) set consensus at 66% and McKenna (9) suggested a 51% level. The level employed depends upon the sample numbers, the aim of the research and resources available (24).

In this study, consensus was considered to have been achieved if more than half of the expert group gave the same score. Types of misconduct for which there was a consensus that it would be likely (score 4) or very likely (score 5) to distort the results of a clinical trial (these two scores being combined for this purpose) were listed with the distribution of opinions on the likelihood that this type of misconduct would actually occur.

2.4 Results

Of the 40 experts invited to take part, 32 agreed to participate in the study, of whom 26 (81%), 27 (84%), and 25 (78%) completed rounds one, two and three respectively. The 26 respondents in round one generated a list of 84 suggestions for the design stage, 93 suggestions for the conduct stage, 88 suggestions for the analysis stage and 85 suggestions for the report stage. Editing and combining similar items reduced the list to 35 suggestions (design), 30 suggestions (conduct), 36 suggestions (analysis) and 42 suggestions (reporting). See Appendix 1.

In the second round, 51 types of scientific misconduct reached the pre-defined level of consensus for being likely or very likely to distort the results of a clinical trial (Table 2.1). All these (51), plus a further (9) types reached the consensus level in round 3.

Table 2.1 Types of misconduct for which consensus was reached on the criterion of likely or very likely to distort the result, with percentages at this level of consensus (round 2)

Types of misconduct	Percentage indicating likely or very likely to distort results
Design	
Failure to use random allocation	88
Failure to specify in the protocol the main outcome measure	66
Inadequate allocation concealment	77
Different follow-up schedules in arms	66
Use of a cross-over where carry-over is expected	76
Intentional use of non-optimum comparison treatment	61
Precision of measurement is avoided in an equivalence trial	66
Inadequate blinding of outcome assessment	69
Inappropriate timing of measurement of treatment effects	51
Conduct	
Tampering with treatment packs so as to un-blind allocation	84
Selective withdrawals on basis of knowledge of allocation	88
Data falsification	85
Data fabrication	85
Treatment recognition in blinded trials	59
Analysis	
Altering analysis methods until find significant result	100
Use of battery of methods of comparison to get the right answer	88
Altering results in knowledge of allocation	100
Excluding patients or results to exaggerate effects or remove adverse events	96
Use of primary outcome measure that was not pre-specified	81
Selecting covariates to bias treatment effect in a particular direction	73
Selective exclusion of "protocol violation outliers"	62
Inappropriate sub-group analyses	81
Claiming equivalence by dint of failure to demonstrate a difference	77
Rely on biased comparisons as the primary analysis	78
Missing data ignored when informative	67
Using a different primary endpoint from that specified in the protocol	80
Post-hoc analysis not admitted	68

Table 2.1 Continued.

Trial stopped for marketing and not scientific reasons	79
Reducing data in a biased fashion	65
Incorrectly imputing values for missing data	56
Sub-group analyses done without interaction tests	65
Failure to account for 'clustering' issues (multi-level)	60
Fail to comply with a pre-specified analysis plan	61
Deviation from intention to treat analysis	62
Ignore data on side-effects	62
Use of inappropriate statistical methods	56
Reporting	
Failure to report unfavourable results	84
Selective reporting of positive results or omission of adverse events data	85
Selective reporting based on p-values	84
Report of sub-group without reference to wide study	85
Pos hoc analyses reported as a main conclusion	77
Negative or detrimental studies not published	79
Over-interpretation of 'significant' findings in small trials	75
Putting undue stress on results from sub-group analysis	76
Selective reporting of (i) sub-groups (ii) outcomes (iii) time points	73
Report of single variable where multiple variables assessed and not reported	60
Failure to report results or long delay in reporting	56
Clinically important effect sizes may be declared to suit results	58
Poor use of figures which mislead / distort results	53
Conclusion drawn that cannot be linked with evidence provided in report	52
Reporting under control of sponsor	54

At the end of the third round, 60 types of scientific misconduct reached the 50% level of consensus for being likely or very likely to distort the results of a clinical trial (Table 2.2). The types of scientific misconduct for which there was consensus that they would be likely or very likely to distort the results and consensus that they would be likely or very likely to occur are shown in Table 2.3. Of the 13 types of misconduct shown in Table 2.3 the most likely to occur was over-interpretation of 'significant' findings in small trials, while selective reporting and inappropriate sub-group analyses were the main themes, these being given as likely to occur by more than three quarters of the respondents.

Table 2.2 Types of misconduct for which consensus was reached on the criterion of likely or very likely to distort the result, with percentages at this level of consensus and the percentage breakdown of respondents' views on the likelihood of occurrence (round 3)

Types of misconduct	Percentage indicating likely or very likely to distort results	Likelihood to occur (%)				
		Very unlikely 1	2	3	4	Very likely 5
Design						
Failure to use random allocation	92	12	68	16	0	4
Failure to specify in the protocol the main outcome measure	88	8	48	28	16	0
Inadequate allocation concealment	84	0	24	48	20	8
Different follow-up schedules in arms	80	8	40	52	0	0
Use of a cross-over where carry-over is expected	79	8	46	46	0	0
Intentional use of non-optimum comparison treatment	76	0	40	44	16	0
Precision of measurement is avoided in an equivalence trial	74	0	30	55	15	0
Inadequate blinding of outcome assessment	72	0	12	72	12	4
Inappropriate timing of measurement of treatment effects	60	4	20	68	8	0
In an equivalence trial, choice of an inappropriate outcome measure	56	0	28	56	16	0
Conduct						
Tampering with treatment packs so as to un-blind allocation	95	17	75	4	4	0
Selective withdrawals on basis of knowledge of allocation	92	8	52	28	12	0
Data falsification	92	64	32	4	0	0
Data fabrication	92	72	24	4	0	0
Treatment recognition in blinded trials	64	4	36	36	24	0
Post-hoc changes in protocol	52	0	20	56	20	4
Analysis						
Altering analysis methods until find significant result	100	4	28	60	8	0
Use of battery of methods of comparison to get the right answer	100	0	24	64	12	0
Altering results in knowledge of allocation	100	76	16	8	0	0
Excluding patients or results to exaggerate effects or remove adverse events	99	17	46	21	16	0
Use of primary outcome measure that was not pre-specified	96	12	48	28	12	0
Selecting covariates to bias treatment effect in a particular direction	96	16	40	32	12	0
Selective exclusion of "protocol violation outliers"	88	0	32	44	24	0
Inappropriate sub-group analyses	88	0	8	28	48	16

Table 2.2 Continued.

Claiming equivalence by dint of failure to demonstrate a difference	88	0	8	42	38	12
Rely on biased comparisons as the primary analysis	87	0	57	30	13	0
Missing data ignored when informative	84	0	20	36	32	12
Using a different primary endpoint from that specified in the protocol	84	16	48	20	16	0
Post-hoc analysis not admitted	83	0	4	37	42	17
Trial stopped for marketing and not scientific reasons	83	0	32	45	14	9
Reducing data in a biased fashion	77	9	43	24	19	4
Incorrectly imputing values for missing data	76	4	36	44	12	4
Sub-group analyses done without interaction tests	75	0	0	25	50	25
Failure to account for 'clustering' issues (multi-level)	72	0	12	44	32	12
Fail to comply with a pre-specified analysis plan	68	0	32	48	16	4
Deviation from intention to treat analysis	68	0	8	60	24	8
Ignore data on side-effects	64	8	40	32	4	16
Fail to specify a reasonable analysis plan in advance	56	0	12	52	20	16
Use of inappropriate statistical methods	56	0	32	48	16	4
Analysis conducted by the sponsor of the trial	54	0	4	42	33	21
Inappropriate analysis for example comparison of survival time by t-test	52	4	32	56	8	0
Reporting						
Failure to report unfavourable results	100	0	8	56	20	16
Selective reporting of positive results or omission of adverse events data	96	0	8	32	24	36
Selective reporting based on p-values	92	0	0	20	64	16
Report of sub-group without reference to wide study	92	0	48	28	24	0
Pos hoc analyses reported as a main conclusion	92	0	32	44	24	0
Negative or detrimental studies not published	88	0	8	24	28	40
Over-interpretation of 'significant' findings in small trials	87	0	0	17	50	33
Putting undue stress on results from sub-group analysis	84	0	4	28	48	20
Selective reporting of (i) sub-groups (ii) outcomes (iii) time points	80	0	4	32	40	24
Report of single variable where multiple variables assessed and not reported	68	0	20	52	20	8
Failure to report results or long delay in reporting	68	0	16	24	24	36
Clinically important effect sizes may be declared to suit results	63	0	12	63	17	8
Poor use of figures which mislead / distort results	60	0	28	56	12	4
Unjustified extrapolation	58	0	17	46	33	4

Table 2.2 Continued.

Selective reporting of outcomes in the abstract	56	0	0	24	44	32
Conclusion drawn that cannot be linked with evidence provided in report	56	4	16	44	20	16
Reporting under control of sponsor	56	0	20	64	8	8
Claim an analysis is by "intention-to-treat" when it is not	52	4	24	48	12	12
Giving incomplete information about analyses with non significant results	52	0	4	40	32	24

Table 2.3 Types of misconduct for which there was consensus (>50%) that they would be likely or very likely to distort the results, and that they would be likely or very likely to occur.

Types of misconduct	Indicating likely or very likely to occur (%)
Over-interpretation of 'significant' findings in small trials	83
Selective reporting based on p-values	80
Selective reporting of outcomes in the abstract	76
Sub-group analyses done without interaction tests	75
Negative or detrimental studies not published	68
Putting undue stress on results from sub-group analysis	68
Inappropriate sub-group analyses	64
Selective reporting of (i) sub-groups (ii) outcomes (iii) time points	64
Selective reporting of positive results or omission of adverse events data	60
Failure to report results or long delay in reporting	60
Post-hoc analysis not admitted	59
Giving incomplete information about analyses with non significant results	56
Analysis conducted by the sponsor of the trial	54

2.5 Conclusion

This study used an expert consensus approach to identify the most important types of scientific misconduct in clinical trials. The most important types of misconduct were considered to be those that occur more commonly and those that distort trial results. Two types of misconduct were established from the Delphi survey in the third round and are inappropriate sub-group analyses and selective reporting of trial results.

The main strength of the Delphi technique is that it optimises the use of group opinion and minimises the bias that can be encountered in face to face group interaction. In this case, all experts offered opinions freely and without any peer pressure from others. The expert panel was chosen because of their knowledge and experience in the conduct of clinical trials.

A limitation of this study was that some of the suggestions elicited in the first round were vague or ambiguous. As a result, it was difficult to accurately exclude duplicates, and so the list that was used in the second and third Delphi rounds was somewhat repetitive. On the other hand, the consistent high ranking of selective reporting and inappropriate sub-group analyses does suggest that these experts' opinions on the most important issues had been accurately identified.

Although there has been considerable attention in the scientific literature to the problems of data fabrication and data falsification, these were absent from our list of the most important types of misconduct, because there was consensus that these problems were very unlikely to occur. The results suggest that selective reporting and the opportunistic use of the play of chance (inappropriate sub-group analyses) are more important considerations in ensuring that patients receive only effective treatments. Indeed, the two problems can be closely related. Multiple post hoc sub-group analysis with selective

reporting might easily result in authors making exaggerated sub-group claims about treatment effectiveness (25).

A publicly accessible inventory of trial protocols that include a clear description of the statistical analysis plan is a potential solution to the problems of selective reporting and sub-group analyses. Such an initiative is already under way, and was given further impetus when the UK NHS joined the worldwide effort to register clinical trials at inception (26-28). Future research will need to assess the extent to which this initiative has been successful.

In revealing these two types of misconduct (inappropriate sub-group analyses and selective reporting of trial results) the Delphi survey supported views that have received much attention in the literature. However, it gives re-affirmation of these and emphasises their potential to distort the results of the trials.

References

1 Linstone HA, Turoff M. The Delphi Method: Technique and Applications. Addison-Wesley Publishing Company; 1975.

2 Johnson D, King M. BASIC forecasting techniques. Butterworths, London 1988.

3 Helmer O. Looking Forward: A Guide to Futures Research. Sage Publications, Beverly Hills 1983.

4 Wang CC, Wang Y, Zhang K, Fang J, Liu W, Luo S, et al. Reproductive health indicators for China's rural areas. *Social Science and Medicine* 2003;57:217-225.

5 Gupta UG, Clarke RE. Theory and Applications of the Delphi Technique: A bibliography (1975-1994). *Technological Forecasting and Social Change* 1996;53:185-211.

6 Robertson HA, MacKinnon NJ. Development of a list of consensus approved clinical indicators of preventable drug-related morbidity in older adults. *Clinical Therapeutics* 2002; 24:1595-1613.

7 Fowles J. Handbook of futures research. Greenwood Press: Connecticut; 1978.

8 Couper MR. The Delphi technique: characteristics and sequence model. *Advances in Nursing Science* 1984; 7:72-77.

9 McKenna HP. The Delphi technique: a worthwhile approach for nursing? *Journal of Advanced Nursing* 1994;19:1221-1225.

10 Goodman CM. The Delphi technique: a critique. *Journal of Advanced Nursing* 1987;12:729-734.

11 Davidson P, Merritt-Gray M, Buchanan J, Noel J. Voices from practice: mental health nurses identify research priorities. *Archives of Psychiatric Nursing* 1997;11:340-345.

12 Lemmer B. Successive surveys of an expert panel: research in decision making with health visitors. *Journal of Advanced Nursing* 1998;27:538-545.

13 Green B, Jones M, Hughes D, Williams A. Applying the Delphi technique in a study of GPs information requirement. *Health and Social Care in the Community* 1999; 7:198-205.

14 Sackman H. Delphi Critique: Lexington Books. Lexington: MA; 1975.

15 Masini E. Why Futures Studies? Grey Seal, London 1993.

16 Webler T, Levine D, Rakel H, Renn O. A Novel Approach to Reducing Uncertainty: The Group Delphi. *Technological Forecasting and Social Change* 1991; 39:253-263.

17 Carley S, Mackway-Jones K, Donnan S. Delphi study into planning for care of children in major incidents. *Archives of Disease in Childhood* 1999; 80:406-409.

18 Butterworth T. Bishop V. Identifying the characteristics of optimum practice: findings from a survey of practice experts in nursing, midwifery and health visiting. *Journal of Advanced Nursing* 1995; 22:24-32.

19 MRC. Clinical trials for tomorrow. London: Medical Research Council; 2003.

20 Erlandson DA, Harris EL, Skipper BL, Allen SD. Doing naturalistic inquiry. A guide to methods. London: Whurr Publishers; 1993.

21 Walker AM, Selfe J. The Delphi method: a useful tool for the allied health researcher. *British Journal of Therapy and Rehabilitation* 1996;3:677-681.

22 Roberts-Davis M, Read SM. Clinical Role Clarification: using the Delphi method to establish similarities and differences between nurse practitioners and clinical nurse specialists. *Journal of Clinical Nursing* 2001; 10:33-43.

23 Boyce W, Gowland C, Russell D, et al. Consensus methodology in development and content validation of a gross performance measure. *Physiotherapy Canada* 1993; 45:94-100.

24 Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing* 2000; 32:1008-1015.

25 Pocock SJ, Assmann SE, Enos LE, and Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trials reporting: current practice and problems. *Statistics in Medicine* 2002; 21:2917-2930.

26 Staessen. JA, Bianchi. G. Registration of trials and protocols. *Lancet* 2003; 362:1009-10.

27 De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *Lancet* 2004; 364: 911-2.

28 Krleža-Jerić K, Chan AW, Dickersin K, Sim I, Grimshaw J, Gluud C. Principles for international registration of protocol information and results from human trials of health related interventions: Ottawa statement (part 1). *BMJ* 2005; 330:956-958.

CHAPTER 3

Selective reporting in clinical trials: analysis of a cohort of trial protocols published by *the Lancet*

3.1 Introduction

In the previous chapter, I reported the results of a Delphi survey that found that inappropriate sub-group analyses and selective reporting of results are believed by the expert panel to be the most common and the most important (most able to distort the trial results) types of scientific misconduct in clinical trials. In the current chapter, I explore these two issues in more detail.

3.1.1 Selective reporting of outcomes

Selective reporting within published studies can occur when multiple outcomes are measured but only some of them are reported (1). It can also occur when investigators conduct many different sub-group analyses but only report the most favourable (2). One can distinguish between unreported, incompletely, and fully reported outcomes (3,4). Incomplete reporting of outcomes causes problems concerning inclusions in meta-analyses.

Selective reporting of results in clinical trial publications is an important source of outcome reporting bias (5-8). Outcome reporting bias can also bias the results of systematic reviews and meta-analyses because the available results are a biased sample of the results from clinical trials (9).

It is often difficult to determine from a clinical trial publication whether additional analyses have been conducted but not reported. Several studies have compared trial

protocols with trial reports to assess the problem of within-study selective reporting (3,4,10). A brief description of the findings from these studies is presented in the next sections.

A small study was undertaken, involving a single local research ethics committee, to examine the extent of the problem of within-study selective reporting in clinical trials (10). The outcome proposed in the original approved trial protocol was compared with the results presented in the subsequent trial report. The primary outcome was stated in only six (40%) of the protocols for the 15 publications obtained. Of these, only four (76%) were consistent with the reports. Eight protocols gave an intended analysis plan, but seven of the publications did not follow this analysis plan. This study suggested that selective reporting might result in considerable bias.

Chan *et al* (3) reviewed a cohort of trial protocols approved by the Scientific-Ethical Committees in Copenhagen and Frederiksberg, Denmark, between 1994 and 1995. The trial protocols were compared with the corresponding published reports to examine the extent and nature of outcome reporting bias. They studied one hundred and two trials with 122 published reports and 3,736 outcomes. They found that 50% of efficacy and 65% of harm outcomes per trial were incompletely reported. 62% of trials had at least one primary outcome that was changed, introduced or omitted. The authors concluded that trial reports are often incomplete, biased or inconsistent with protocols. Moreover, published articles may be unreliable and overestimate the effect of the treatment.

A similar study (4) comparing trial protocols approved by the Canadian Institute of Health Research from 1990 to 1998 with subsequent reports was conducted to determine whether outcome reporting bias would also be present in government-funded trials that had been subjected to rigorous peer review. Forty-eight trials with 68 publications and 1402 outcomes were identified. A median of 31% of efficacy outcomes and 59% harm

outcomes per trial were incompletely reported. Primary outcomes differed between protocols and reports in 40% of trials. The authors concluded that selective reporting of outcomes occurs even in high-quality government-funded trials.

3.1.2 Sub-group analyses

Establishing the overall effect of a treatment among a group of patients is the main aim of a clinical trial. However, patients recruited for a clinical trial are not a homogeneous sample. Their responses to treatment may vary, such that the treatment may be more or less effective in particular sub-groups of patients. If there are specific sub-groups of patients for which there is evidence that a treatment may be more or less effective than is indicated by the overall effect in the trial as a whole, then there is a responsibility to explore such sub-group effects.

The problems associated with sub-group analyses have been discussed by several authors (8,11,12). The play of chance can produce spurious results in sub-group analyses in two ways. First, there is an increased statistical likelihood of a false positive finding when many sub-groups are examined (13). By definition, testing at the 5% level of significance will erroneously report statistically significant differences in about 5% of the tests performed (so-called false-positive results), i.e. one false positive out of twenty independent outcomes are studied. Secondly, there is a risk of false negative results, the failure to detect a true difference in effect in a sub-group. Power calculations for the trial are usually based on the detection of an overall pre-specified treatment effect in all trial participants. Sub-group analysis will have reduced power to detect differential effects. Therefore, sub-group analyses are problematic both in terms of false positive and false negative results.

The recommended statistical approach is the use of formal tests of interaction, which directly examine the treatment difference between sub-groups (14). Although interaction tests go some way to addressing concerns about false positive conclusions, a statistical test for interaction is unlikely to be definitive because of low statistical power. The sample sizes within sub-groups are often small and interaction tests require large sample sizes (15). If there is a strong belief that particular sub-groups may show stronger treatment effects, but the group interaction tests are not significant, further evidence from other studies with similar study populations is generally required to confirm the sub-group findings.

Because of the hazards of inappropriate sub-group analyses, all clinical trials should have a pre-defined statistical analysis plan that includes full details of all sub-group analyses to be conducted, with supporting evidence to justify the sub-group analyses. The plan should specify in detail what analyses would be conducted and how sub-groups would be defined. Comparison of the statistical analysis plan with the trial report may help to avoid the problem of authors failing to state in the final report how many less exciting sub-group analyses were conducted but not reported. In addition, sample size estimates should have any sub-group analyses in mind, and the interpretation of the significance level should consider multiple comparisons (e.g. the Bonferroni correction).

3.1.3 Publication of trial protocols

A number of guidelines (16,17) have been developed with the aim of improving the quality of design, conduct and reporting of clinical trials. Of particular relevance is the registration of the trials and publication of the trial protocols (18-23) to deter bias and to avoid duplication research. Protocol publication allows comparison between what was originally planned and what was actually carried out, and reduces the potential for 'data dredging'. Deviation from protocols should be identified and the reasons for it need to be

clearly presented (7). In this regard, the BMJ and *the Lancet* have stipulated that authors must submit trial protocols at the time of manuscript submission. They will not send trial manuscripts for external review unless they are accompanied by the protocols (24,25).

This chapter reports an investigation of the extent of selective reporting and inappropriate sub-group analyses in a cohort of clinical trial protocols that had been peer reviewed and accepted by *the Lancet*. My objective is to: (1) determine the prevalence of selective reporting of primary and secondary outcomes and sub-group analyses in published reports of randomised trials, (2) examine the consistency between the protocols and published reports, and (3) assess the adequacy of pre-specification of sub-group analysis in trial protocols.

3.2 Methods

The Lancet began reviewing protocols for clinical trials in 1997. Protocols deemed to be of general medical interest are sent out for clinical and statistical review. By July 2004, *the Lancet* had accepted fifty-six protocols, summaries of which were published on their web site (26). Their aim was threefold: to encourage good principles in the design of clinical research, to publicise a list of "accepted" protocols, and to make a provisional commitment to publication of the main clinical endpoints of the study. The instructions to authors ask that all protocols should address any planned sub-group analyses. Using information from *the Lancet* web site, I contacted the authors of all published protocols and asked if they would either provide a copy of the full trial protocol or give permission for *the Lancet* to release the full protocol for inspection. The research ethics committee of the London School of Hygiene & Tropical Medicine approved this study (Appendix 2).

I first extracted data on all primary and secondary outcomes and sub-group analyses as specified in the trial protocols. If a protocol specified primary and secondary aims or objectives, then I took these to indicate primary and secondary outcomes. I also extracted data on year of submission of the protocol, source of funding, number of randomized groups, and the number of study subjects and centres.

I then searched PubMed for any published trial reports arising from this cohort of protocols using investigator name and *the Lancet,* or using investigator name and terms for the intervention in order to find publications in other journals (final search, February 2005). Using the contact details from the protocols, I wrote an e-mail to the authors of the protocols and asked them to provide details of all relevant trial reports.

Having located the relevant trial reports, I extracted data on primary and secondary outcomes and sub-group analyses and compared these data with the data from the trial protocols. I collected data on all outcomes that had been listed in the trial protocol but were absent from the "Results" section of the published reports. Similarly, I collected data on all primary and secondary outcomes and sub-groups that were present in the "Results" sections of published reports, but were not mentioned in the trial protocols. Any discrepancies between protocols and trials were tabulated (the outcome being recorded as absent from either the trial protocol or the reports). I recorded whether the primary and secondary outcomes and sub-group analyses in the publications were significant or not at the $p \leq 0.05$ level of significance. I also collected data on outcomes specified as the primary outcome in the protocol but reported as a secondary outcome in the trial report and outcomes specified as the primary outcomes in the protocol but not explicitly defined as such in the "Results" section.

For sub-group analyses, I determined the number of sub-group factors (for example sub-group analyses may use age as a factor). I also documented whether any hypothesis had

been given to justify the sub-group analyses, and whether the precise definition of each sub-group and the total number of possible sub-group analyses was clearly specified. I documented the statistical method used to assess sub-group effects including the use of interaction tests and any correction for multiple statistical testing. I also assessed whether sample size calculations were conducted for the sub-group analysis.

3.3 Results

From the 56 protocols accepted by *the Lancet*, I excluded four non-randomized trials. Of the remaining 52, I obtained permission to use 48 (92%), but no reply was received from the remaining four investigators. *The Lancet* provided me with the latest version of each protocol, so that I could take into account any amendments. Of the 48 trials, the principal investigators of 18 (38%) provided additional information on where their trial was published. Two stated that their trial had not been published. After contacting investigators and searching the database, I found one or more publications for 30 (63%) of the 48 included trials.

The 48 trial protocols examined had the following characteristics: 40 had two randomized groups, 4 had three groups, and 4 had four groups. Sample sizes ranged from 54 to 20,000 patients with a median of 807. There were 40 multi-centre trials. The number of primary outcomes varied substantially. 27 had one predefined primary outcome, 15 had two primary outcomes, and 6 had three or more primary outcomes. The number of secondary outcomes varied between none and eighteen.

Twenty-four (50%) of the protocols mentioned that sub-group analyses would be undertaken. Five confined sub-group attention to one factor (such as age or sex), but 19 mentioned more than one. Only three protocols gave the hypothesis motivating the selection of sub-groups. None specified the actual number of sub-groups. In 11

protocols, information on levels for at least one of the factors was not given, so that the number of sub-groups could not be calculated. Three protocols mentioned correction for multiple statistical testing when several statistical tests were to be performed simultaneously (sub-group or multiple primary outcomes). Four reported that they accounted for sub-group analyses in their sample size estimates. Four mentioned the use of small p-values for multiple statistical testing. Five mentioned the use of statistical tests for interaction in evaluating sub-group analyses.

The thirty trial protocols whose papers had been published, examined and had the following characteristics: 25 had two randomized groups, 2 had three groups, and 3 had four groups. Sample sizes ranged from 113 to 20,000 patients with a median of 807 (10^{th}-90^{th} percentile range 202-13,800). Most of the trials (n=24, 80%) involved multiple study centres. The year of the preparation of eight protocols was not available.

There were 42 published reports arising from the 30 published trials (27-68). Of these, 27 (64%) were published in *The Lancet* and the remaining 15 in specialty journals. Twenty-nine (97%) of the trials were funded by non industrial sources, e.g. MRC or NHS. One trial did not report the source of funding. The year of submission and publication for both protocols and trial reports is shown in Table 3.1 below. Eight protocols did not mention the date of submission. The median interval between publication of the protocol and publication of results was 5 years (10^{th}-90^{th} percentile range 3.0-8.4). Sample sizes ranged from 108 to 19,025 patients with a median of 697 (10^{th}-90^{th} percentile range 183-10,991).

Comparison of trial protocols with reports

The number of primary outcomes in the protocols varied substantially. 17 had one pre-defined primary outcome, 7 had two, 4 had three, and 2 had four primary outcomes. In

summary, a total of 51 primary outcomes were specified with a median of 1 per trial (10^{th}-90^{th} percentile range 1-3). The number of secondary outcomes varied between none and seventeen; a total of 133 secondary outcomes were defined, with a median of 4 outcomes per trial (10^{th}-90^{th} percentile range 0-10). Fourteen trials mentioned sub-group analysis.

In the published reports, there were 60 primary outcomes reported, a median of 2 per trial (10^{th}-90^{th} percentile range 1-4). There were 94 secondary outcomes, a median of 2 per trial (10^{th}-90^{th} percentile range 0-8). Twenty-two trials examined sub-group analyses. In nine of the trials, the sub-group analyses had not been pre-specified in the protocols. In one of the fourteen trials mentioned above, the sub-group analysis was not reported in the publication (Table 3.2).

Discrepancies in reporting primary outcomes

Of the 30 trials, in eleven (37%) there were major discrepancies between the protocols and the reports regarding primary outcomes (an unreported primary outcome, a new primary outcome or a protocol primary outcome becoming secondary outcome in the published report) (see Table 3.3). Five trials had an unreported primary outcome. Seven trials introduced a new primary outcome. In two trials the protocol primary outcome was reported as secondary in the published report. None of these trials gave any reasons for including or omitting outcomes or changing their status.

Discrepancies in reporting secondary outcomes

Twenty-six trials (87%) had at least one unreported secondary outcome or at least one new secondary outcome. Twenty-two trials had one or more unreported secondary outcomes (median of 2 unreported outcomes 10^{th}-90^{th} percentile range 1-6). Eleven trials introduced one or more new secondary outcomes (median of 1 new outcome, 10^{th}-90^{th} percentile range 1-6) (Table 3.4).

Sub-group analyses in the 30 protocols

Fourteen protocols pre-specified one or more sub-group analysis and sixteen did not (Figure 3.1). In the 14 protocols that mentioned that sub-group analyses would be undertaken, three confined sub-group attention to one factor, but 11 mentioned more than one factor. Only one protocol gave the hypothesis motivating the selection of sub-groups. None specified the total number of sub-groups. In seven protocols, information on levels for at least one of the factors was not given, so that the number of sub-groups could not be estimated. Three protocols mentioned correction for multiple statistical testing when several statistical tests were to be performed simultaneously (sub-group or multiple primary outcomes). Two reported that they accounted for sub-group analyses in their sample size estimates. Five mentioned the use of statistical tests for interaction in evaluating sub-group analyses.

Sub-group analyses conducted but not specified in the protocol

Among the trials with no specified sub-group analyses in the protocol, sub-group analyses were conducted in 9/16 (56%). Four of these reported that the sub-group analyses had been pre-specified, although no evidence could be found for this in the protocol. None gave the hypothesis motivating the sub-group analyses. Two trials reported the results of interaction tests. None adjusted the significance level for the sub-group analyses or the sample size for sub-groups.

The main effect and the sub-group effects were the same (non-significant, significant) in 22% (2/9) and 11% (1/9) of the trials respectively. In two trials (22%), the overall results were non-significant, but at least one sub-group result was significant. In four trials (44%), the overall results were significant but at least one sub-group result was non-significant.

Sub-group analyses different to those specified in the protocol or reported without using appropriate statistical methods

Among the fourteen trials where sub-group analyses were mentioned in the protocol, seven trials (50%) had at least one sub-group unreported. Two trials (14%) reported more sub-groups than specified in the protocol. In two trials (14%), the sub-group definitions in the report were different to those in the protocol. The total number of sub-groups in the papers was the sum of the number of levels of the defining factors in all of the trial reports. There were no instances of sub-groups based on combination of factors (eg. age by sex). One trial reported the hypothesis for selected sub-groups. Six trials reported the results of interaction tests. In no trial was the significance level adjusted for multiple statistical testing, or the sample size increased for sub-group analyses.

Table 3.1 Year of publication of protocols and trial reports (n=22)

	Publication year of the main reports					
	2000	2001	2002	2003	2004	2005
Submission year of the protocols						
1995	1	1			1	
1996		2	1			1
1997		1	1	1	3	
1998			1	1		1
1999				1	1	1
2000				1		
2001				1	1	

Table 3.2 Data on primary and secondary outcomes and sub-group analysis extracted from protocols and corresponding published articles

Hormonal replacement therapy versus best symptomatic treatment without hormones for women with previous breast cancer

Trial identification	Protocol	Reports	Significance level
1	**Primary**		
	Risk of breast cancer recurrence	New breast cancer events *	S
	Secondary		
	Quality of life	Absent	
	Risk of breast cancer death	Absent	
	Myocardial infarction	Absent	
	Cerebral stroke	Absent	
	Fracture of the spine and hip	Absent	
	Sub-group		**No. of groups with significant results**
	Not mentioned	Hormone-receptor status (2 groups)	1
		Tamoxifen or not (2 groups)	1
		Hormone replacement therapy before diagnosis or not (2 groups)	1

* Reported as a primary outcome
S: Significant
NS: Not Significant

Open versus laparoscopic-assisted surgery in patients with colorectal cancer

Trial identification	Protocol	Reports	Significance level
2	**Primary**		
	Absent	Proportion of Dukes's C2 tumours	NS
	Absent	Cancers of the colon or of the rectum	NS
	Circumferential, longitudinal and high tie mesenteric resection margins	Circumferential and longitudinal resection margins	NS
	30-day operative mortality	Absent	
	Absent	Hospital mortality	NS
	Secondary		
	Complication rates	Intra-operative complication	NS
		Post-operative complication	NS
	Quality of life and cost effectiveness	Quality of life during 3 months after surgery	NS
	Transfusion requirements	Transfusion requirements	NS
	Port-site and wound site recurrences	Absent	
	Loco-regional, anastigmatic and distant metastases	Absent	
	Sub-group		
	Not mentioned	None	

NS: Not Significant

Low dose aspirin versus placebo for patients who undergoing surgery for hip fracture

Trial identification	Protocol	Reports	Significance level
3	**Primary**		
	All vascular deaths	All vascular deaths	NS
	Major non-fatal vascular events (pulmonary embolism, myocardial infarction or stroke) in hospital before day 35	Deep-vein thrombosis	S
		Pulmonary embolism	S
		venous thromboembolism	S
		Myocardial infarction	NS
		Stroke	NS
	Bleeding complications (haematoma requiring evacuation, post-operative bleed requiring transfusion) in hospital before day 35	Haematoma requiring evacuation	NS
		Bleeding after surgery with transfusions	NS
	Secondary		
	Vascular mortality at 6 month and one year	Absent	
	All cause mortality at 35 day and later	Absent	
	Sub-group		**No. of groups with significant results**
	Not mentioned	Age (3 groups)	1
		Sex (2 groups)	2

		Fracture site (2 groups)	2
		Previous used of non-steroidal anti-inflammatory drugs (2 groups)	2
		Time of first dose (2 groups)	1
		Heparin (3 groups)	2
		TED stockings (2 groups)	1
		Surgical procedures (2 groups)	1
		Regional anaesthesia (2 groups)	1

NS: Not Significant
S: Significant

Perindopril-based blood-pressure-lowering regimen versus placebo in patients with previous stroke or transient ischaemic attack

Trial identification	Protocol	Reports	Significance level
4	**Primary**		
	Stroke	Stroke	S
	Absent	*Fatal or disabling*	S
	Absent	*Not fatal or disabling*	S
	Absent	*Ischaemic stroke*	S
	Cerebral haemorrhage	*Cerebral haemorrhage*	S
	Stroke of unknown type	*Stroke of unknown type*	NS
	Secondary		
	Fatal and total cardiovascular events	vascular death	NS
	Cognitive function and dementia	Cognitive function	NS
	Disability and dependency	Absent	
	Other death and hospitalisations	Total deaths	NS
	Cerebrovascular events	Absent	
	Sub-group		**Interaction†**
	Not mentioned	Qualifying cerebrovascular event (2 groups)	NS
		Time between the qualifying event and enrolment (2 groups)	NS

		Geographic region of residence (2 groups)	NS
		Separate estimate of treatment effects among participants for whom combination therapy was planned at randomisation and those for whom single drug therapy was planned (2 groups)	S
		Hypertensive (2 groups)	S

NS: Not Significant
S: Significant
† The significant level from the interaction test

Routine clinical diagnosis versus routine clinical diagnosis plus genetic testing for familial hypercholesterolaemia within a previously aware population

Trial identification	Protocol	Reports	Significance level
5	**Primary**		
	Impact of familial hypercholesterolaemia upon risk reduction behaviours (smoking, diet, taking medication and exercise)	Risk reduction behaviours§	NS
	Absent	Perceived familial hypercholesterolaemia	S
	Absent	Cholesterol	S
	Absent	Heart disease	S
	Secondary		
	Emotional state	Emotional state	S
	Illness cognitions	Illness and treatment perceptions	S
	Absent	Perceptions of the diagnosis	S
	Sub-group		
	Not mentioned	None	

NS: Not Significant
S: Significant
§ Reported as secondary

Epidural versus intra-operative anaesthesia and analgesia and post-operative analgesia in patients with major abdominal surgery

Trial identification	Protocol	Reports	Significance level
6	**Primary**		
	Mortality in hospital or within 30 days	Mortality in hospital or within 30 days	NS
	Major morbidity:		
	Respiratory failure	*Respiratory failure*	S
	Cardiovascular events	*Cardiovascular events*	NS
	Renal failure	*Renal failure*	NS
	Gastrointestinal failure	*Gastrointestinal failure*	NS
	Hepatic failure	*Hepatic failure*	NS
	Haematological failure	*Haematological failure*	NS
	Infection, pneumonia, sepsis	*Inflammation/ sepsis*	NS
	Secondary		
	Length of time in ICU	Absent	
	Cost of care	Absent	
	Sub-group		
	Not mentioned	None	

NS: Not Significant
S: Significant

Neurosurgical clipping versus endovascular coiling in patients with ruptured intracranial aneurysm

Trial identification	Protocol	Reports	Significance level
7	**Primary**		
	Reduce the proportion of patients with moderate or poor outcome (defined by Rankin grade 3-6) by 25% at one year	Death or dependence at one year (defined by Rankin grade 3-6) *	S
	Secondary		
	Preventing re-bleeding	The frequency of non-procedural re-bleeding is shown	
	Quality of life at one year (Euroqol measure)	Absent	
	Cost effectiveness	Absent	
	Improve the neuropsychological at 1 year	Absent	
	Absent	Seizure occurrence	S
	Sub-group		**Interaction†**
	Not mentioned	Age (5 groups)	S
		World Federation of Neurosurgical Societies (WFNS) (2 groups)	S
		Amount of blood on CT scan (Fischer grade) (2 groups)	NS
		Lumen size (3 groups)	NS
		Aneurysm location (4 groups)	S

* Reported as a primary outcome
NS: Not Significant
S: Significant
† The significant level from the interaction test

Cannabis extract, Δ^9-tetrahydrocannabinol or placebo in patients with stable multiple sclerosis and muscle spasticity

Trial identification	Protocol	Reports	Significance level
8	**Primary**		
	Change in Ashworth score	Change in Ashworth score*	NS
	Secondary		
	Rivermead mortality index	Rivermead mortality index	NS
	Time for 10m walk	Time for 10m walk	S
	Four self completion questionnaires	Absent	
	The UK neurological disability score	The UK neurological disability score	NS
	The Barthel index	The Barthel index	NS
	The general health questionnaire	The general health questionnaire	NS
	kings health questionnaire	Absent	
	Absent	Symptom improvement:	
		Bladder	*NS*
		Pain	*S*
		Tremor	*=.05*
		Spasticity	*S*
	Sub-group		
	Not mentioned	Ambulatory or non-ambulatory patients	S

* Reported as a primary outcome
NS: Not Significant
S: Significant

Low-dose versus high-dose interferon-αn1 (IFN) in patients with chronic myeloid leukaemia

Trial identification	Protocol	Reports	Significance level
9	**Primary**		
	Duration of survival	Duration of survival	NS
	Secondary		
	Toxicity assessed by percentage of patients requiring dose reduction or abandoning therapy	Actuarial risk of abandoning IFN for any reason	NS
	Absent	Actuarial risk of abandoning IFN for any reason other than transplantation and disease	S
	Haematologic and cytogenetic response at 6 monthly interval	Haematologic and cytogenetic response at 6 monthly interval	NS
	Absent	Quality of life assessed by EORTC questionnaire	NS
	Sub-group		
	Not mentioned	Sokal risk group (3 groups)	NS
		European risk group (4 groups)	NS

NS: Not Significant
S: Significant

Colposcopy versus cytology and human papillomavirus (HPV) test in women with positive for high-risk types of HPV

Trial identification	Protocol	Reports	Significance level
10	**Primary**		
	Detection rate and positive predictive value	Detection rate and positive predictive value	S
	Secondary		
	Reducing false positive results	Absent	
	The ability of Thin Prep slides to improve the detection rate and positive predictive value	Absent	
	Examine the appropriate recall interval for women with HPV negative borderline smears, or HPV positive but not cytologically negative smears	Absent	
	Absent	Low-grade lesions	S
	Sub-group		
	Not mentioned	None	

S: Significant

Early surgery versus initial conservative treatment in patients with spontaneous supratentorial intracerebral haematomas

Trial identification	Protocol	Reports	Significance level
11	**Primary**		
	Glasgow Outcome Scale	Glasgow Outcome Scale	NS
	Secondary		
	Rankin scale	Rankin scale	NS
	The Barthel index	The Barthel index	NS
	Days in hospital	Absent	
	EuroQol	Absent	
	Sub-group		
	Not mentioned	Age (2 groups)	NS
		GCS (3 groups)	NS
		Side of haematoma (2 groups)	NS
		Site of haematoma (2 groups)	NS
		Haematoma volume (2 groups)	NS
		Depth from cortical surface (2 groups)	S
		Intended methods of evacuation (2 groups)	NS
		Deficit of affected arm (2 groups)	NS
		Deficit of affected leg (2 groups)	NS

		Deficit of speech (3 groups)	NS
		Any thrombolytic or anticoagulant treatment (2 groups)	NS
		Country (12 groups)	NS

NS: Not Significant
S: Significant

Tacrolimus versus microemulsified cyclosporin in liver transplantation patients

Trial identification	Protocol	Reports	Significance level
12	**Primary**		
	Time to reach re-transplantation	Time to reach re-transplantation*	S
	Death	Death*	S
	Treatment failure for immunological reasons	Treatment failure for immunological reasons*	S
	Absent	Death or re-transplantation	S
	Absent	Death or re-transplantation or treatment failure for immunological reasons	S
	Secondary		
	Absent	Hepatic artery thrombosis	S
	Survival	Survival	S
	Acute rejection	Acute rejection	NS
	Steroid resistant cellular rejection	Steroid resistant cellular rejection	NS
	Chronic rejection	Chronic rejection	NS
	Total supplementary steroid usage	Absent	
	Withdrawals from the original immunosuppressive regimen	Absent	
	A change from protocol immunosuppression	Absent	
	Absent	Renal dysfunction	NS
	Absent	Diabetes after 3 months	S

	Sub-group		
	Not mentioned	Elective or emergency treatment	S
		Treatment started within 6h of skin closure or not	S

* Report as a primary outcome
NS: Not Significant
S: Significant

Written feedback followed by course, course alone, written feedback alone orcontrol for oncologists

Trial identification	Protocol	Reports	Significance level
13	**Primary**		
	Multiple and leading questions	Leading questions	NS
	Checking patients' understanding	Checks understanding	NS
	Provision of appropriate reassurance, empathy	Expressions of empathy	S
	Better detection of psychological distress	Absent	
	Secondary		
	Facilitative utterances	Absent	
	Use of Jargon	Absent	
	Appropriate use of open-ended questions	Focused and open questions	S
	Improved self-critique ability	Absent	
	Identification of patients' cues	Appropriate responses to patients' cues	NS
	Sub-group		
	Not mentioned	None	

NS: Not Significant
S: Significant

Immediate delivery or deferred delivery until the obstetrician was no longer uncertain for infant in European countries

Trial identification	Protocol	Reports	Significance level
14	**Primary**		
	Survival to hospital discharge	Death at two years*	NS
	Griffith's full scale at two years	Griffith's full scale at two years*	NS
	Secondary		
	None	None	
	Sub-group		
	Not mentioned	None	

* Reported as a primary outcome
NS: Not Significant

Influenza vaccination versus placebo in children with asthma

Trial identification	Protocol	Reports	Significance level
15	**Primary**		
	Number,	Number*,	NS
	duration,	duration*,	NS
	and severity of asthma exacerbations associated with virologically proven influenza infection	and severity of asthma exacerbations associated with virologically proven influenza infection*	NS
	Secondary		
	Number, duration and severity of asthma exacerbations associated with other in the throat swab detected causative agents	duration of all asthma exacerbations irrespective of causative agents	S
	Proportion of days without symptoms of asthma	Proportion of days with asthma symptoms	NS
	Proportion of days without symptoms of upper respiratory tract infection	Proportion of days with upper respiratory tract symptoms	NS
	Proportion of days without respiratory symptoms	Proportion of days with respiratory symptoms	NS
	Rising of antibody-titre against influenza	Rising of antibody-titre against influenza	S
	Quality of life effects in the first week of the reported episodes	Absent	

	Quality of life in the observed periods of upper or lower respiratory tract symptoms	Absent	
	Number of serologically proven influenza infections	Number of serologically proven influenza infections	S
	Number of influenza infections not reported	Absent	
	Cost-effectiveness ratio	Absent	
	Fall in FEV_1, PEF, FVC, MEF_{50}, measured within 72 hours of the reported period	Absent	
	Fall in FEV_1, PEF, FVC, MEF_{50}, measured in the period of lower respiratory tract symptom	Absent	
	Reported use of asthma medication	Use of asthma medication	NS
	Adverse effects of the vaccination, measured day 1-7 after vaccination	Adverse effects	S
	Consultations for asthma at specialist or GP, admittance to hospital for asthma	Consultations for asthma at specialist or GP, admittance to hospital for asthma	NS
	Consultations for respiratory tract infections by specialist or GP, admittance to hospital for respiratory tract infections	Absent	
	Other detected viruses in lower and upper respiratory tract infections	Absent	
	Sub-group		
	Not mentioned	Vaccination history (2 groups)	NS

* Reported as a primary outcome
NS: Not Significant
S: Significant

The combination of an educational package versus the use of mite allergen-impermeable mattress encasings in children at increased atopic

Trial identification	Protocol	Reports	Significance level
16	**Primary** The overall sensitization rate against house dust mite (HDM) and manifestation of atopic disease	Sensitization against HDM and manifestation of atopic disease	S
	Secondary Efficiency and cost effectiveness under the different health plans of five European countries	Absent	
	Sub-group Not mentioned	None	

S: Significant

Trial identification	Protocol	Reports	Significance level
17	**Primary**		
	Severe postpartum haemorrhage (>1000 mL)	Blood loss ≥ 1000 mL*	S
	The need for additional uterotonics	Use of additional ulterotonics*	S
	Secondary		
	Postpartum haemorrhage (> 500 mL)	Blood loss ≥500 mL	<.05
	Blood transfusion	Need for blood transfusion	NS
	Manual removal of placenta	Manual removal of placenta	NS
	Delay postpartum haemorrhage	Delayed postpartum haemorrhage	NS
	Other measures of severe postpartum Haemorrhage (bimanual compression, Exploration under general anaesthetic hysterectomy, intensive care unit admission)	Bimanual compression	NS
		Exploration under general anaesthetic	NS
		Hysterectomy	NS
		Intensive care unit admission	NS
	Absent	Maternal death	NS
	Sub-group		
	Parity	Parity	NS
	Oxytocin before delivery	Oxytocin or prostaglandin use before delivery	NS
	Epidural analgesia	Absent	

* Reported as a primary outcome
NS: Not Significant
S: Significant

Peer-led versus usual teacher-led sex education for pupils in England schools

Trial identification	Protocol	Reports	Significance level
18	**Primary**		
	Absent	Timing of sex education was right	S
	Unprotected sexual intercourse before the age of 16	Proportion reporting unprotected first intercourse	NS
	Regretted sexual relationships	Regretted first intercourse	NS
	Incidence of termination of pregnancy by age 19-20	Absent	
	Secondary		
	Pupil satisfaction with school sex education	Absent	
	Ability to access sexual health services	Absent	
	Condom use at first and last intercourse	Condom used	S
	Appropriate use of contraception	Absent	
	Reasons for continuing, extending or changing the method of sex education used during the study	Absent	
	Staff attitudes to future provision of sex education	Absent	
	Chlamydia and herpes genital infections	Absent	
	Absent	Saying no to unwanted sex	S
	Sub-group		
	Risk school (low, medium, high)	Absent	

NS: Not Significant
S: Significant

Inhaled nitric oxide versus ventilatory support with out inhaled nitric oxide for preterm infants with severe respiratory failure

Trial identification	Protocol	Reports	Significance level
19	**Primary**		
	Death or severe disability at 1 year of age	Death or severe disability at 1 year*	NS
	Death before discharge from hospital	Absent	
	Secondary		
	Referral for extracorporeal membrane oxygenation	Absent	
	Length of stay in hospital	Duration of time in hospital	NS
	Length of time on supplemental oxygen	Duration of time on supplemental oxygen	NS
	Length of time on ventilator	Duration of time on ventilator	NS
	Pneumothorax or other pulmonary airleak	Pneumothorax or other pulmonary airleak	
	Pulmonary haemorrhage	Pulmonary haemorrhage	
	Major cerebral abnormality	Major cerebral abnormality	
	Necrotising enterocolitis	Absent	
	Treatment of retinopathy of prematurity	Retinopathy of prematurity	
	Infection suspected or confirmed	Infection suspected or confirmed	
	Age at which oral feeding is established	Absent	
	Sub-group		**Interaction†**
	Gestational age (<34 weeks, ≥34 weeks)	Absent	

	Respiratory distress	Death or severe disability:	
	Postnatal age	*Postnatal age:(≤3d,>3d)*	*NS*
	Severity of respiratory disease	*Diagnosis:(Acute, chronic, other)*	*NS*
		Severity (OI≤30, OI>30)	*NS*
		Death or supplemental oxygen at day of delivery:	
		Postnatal age:(≤3d,>3d)	*S*
		Diagnosis:(Acute, chronic, other)	*NS*
		Severity (OI≤30, OI>30)	*NS*

* Reported as a primary outcome
NS: Not Significant
S: Significant
† The significant level from the interaction test

Endovascular versus surgical treatment in patients with carotid stenosis in the carotid and vertebral artery transluminal angioplasty

Trial identification	Protocol	Reports	Significance level
20	**Primary**		
	Transient ischaemic attacks	Absent	
	Stroke and death occurring during and within 30 days	Stroke and death within 30 days	NS
	Disabling stroke or death	Disabling stroke	NS
		Non- disabling stroke	NS
		Death or disabling stroke	NS
		Death or any stroke	NS
	Secondary		
	Use of guide wires	Absent	
	Identify risk factors, eg calcified plaques	Absent	
	Recurrence rate of stenosis	Recurrent events	NS
	Absent	Cranial nerve palsy	S
	Absent	Peripheral nerve palsy	NS
	Absent	Haematoma (requiring surgery or extending hospital stay)	S
	Absent	Myocardial infarction (non-fatal)	NS
	Absent	Pulmonary embolus	NS

	Sub-group		
	Eligible or ineligible for surgery	Absent	
	Symptomatic or asymptomatic	Absent	
	Degrees of stenosis	Severity of stenosis (2 groups)	NS

NS: Not Significant
S: Significant

Magnesium sulphate versus placebo for with pre-eclampsia and their babies

Trial identification	Protocol	Reports	Significance level
21	**Primary**		
	Eclampsia	Eclampsia	S
	Death of the baby before discharge from hospital	Baby dying	NS
	Maternal deaths	Maternal mortality	NS
	Serious maternal morbidity	Maternal morbidity§	NS
	Secondary		
	Use of maternal health service	Use of hospital resources	NS
	Complications of labour and delivery	Complications of labour	S
	Neonatal morbidity	Neonatal morbidity	NS
	Use of neonatal health service	Absent	
	Toxicity	Absent	
	Side effects of magnesium sulphate	Side effects	NS
	Absent	Caesarean section	S
	Sub-group		
	Severity of pre-eclampsia (4 groups)	Severe pre-eclampsia (2 groups)	S

	Delivered or not	Randomised before or after delivery	S
	Anticonvulsant or not	Anticonvulsant or not	S
	Imminent eclamsia or not	Imminent eclamsia or not	S
	Gestational age <33 weeks or not	Gestational age <33 weeks or not	NS

NS: Not Significant
S: Significant
§ Reported as secondary

Corticosteroid versus placebo in adults with significant head injury

Trial identification	Protocol	Reports	Significance level
22	**Primary**		
	Death from any cause within 2 weeks of injury	Death from any cause within 2 weeks of injury	S
	Death or dependence at 6 months	Death or dependence at 6 months	S
	Secondary		
	None	None	
	Sub-group		**Interaction†**
	Time of injury to the initiation of treatment	Time of injury to the initiation of treatment (3 groups)	NS
	Severity of head injury	Severity of head injury (3 groups)	NS
	Absent	CT scan (8 groups)	1

NS: Not Significant
S: Significant

Oral prednisolone versus placebo in children with viral wheeze

Trial identification	Protocol	Reports	Significance level
23	**Primary**		
	Mean day-time and night-time lower respiratory tract symptoms scores	Mean day-time and night-time lower respiratory tract symptoms scores	NS
	Secondary		
	Metered dose inhaler β2-agonist use	Absent	
	Nebulised β2-agonist use	Absent	
	Need for hospital admission	Admission to hospital	NS
	Need for stopping trial medication	Need for stopping trial medication	NS
	Sub-group		**Interaction†**
	High or low eosinophil protein X (EPX)	High or low EPX	NS
	Used of Urinary EPX or not	Absent	

NS: Not Significant
† The significant level from the interaction test

Glasses and patching compared with glasses or no treatment for unilateral visual impairment detected at preschool vision screening

Trial identification	Protocol	Reports	Significance level
24	**Primary**		
	LogMAR visual acuity	LogMAR visual acuity	NS
	Secondary		
	None	None	
	Sub-group		**Interaction†**
	Centres	Centres	NS
	Different initial acuities	Initial acuities (2 groups)	S

NS: Not Significant
S: Significant
† The significant level from the interaction test

Surgery and compression versus compression in patients with chronic venous ulceration

Trial identification	Protocol	Reports	Significance level
25	**Primary**		
	Ulcer healing rate	Ulcer healing rate	NS
	Ulcer recurrence rate	Ulcer recurrence rate	S
	Secondary		
	Cost effectiveness	Absent	
	Quality of life changes	Absent	
	Sub-group		
	Deep venous reflux	Venous reflux patterns (3 groups)	NS

NS: Not Significant
S: Significant

Erythromycin, co-amoxiclav, erythromycin and co-amoxiclav, or placebo in women with preterm, prelabour rupture of fetal membranes

Trial identification	Protocol	Reports	Significance level
26	**Primary**		
	Absent	Delivery within 48 h	S
	Absent	Delivery within 7 days	NS
	Death or major chronic lung disease or cerebral abnormality on ultrasound in the baby before discharge from hospital	Deaths	NS
	Secondary		
	Gestation at birth	Gestational age at delivery	
		<37 weeks	*NS*
		32-36 weeks	*NS*
	Birth weight	Birth weight (g)	
		<2500	*NS*
		<1500	*NS*
	Respiratory distress syndrome	Absent	
	Infection in baby confirmed by blood culture	Positive blood culture	S
	Treatment with surfactant	Treatment with exogenous surfactant	=.05
	Length of time in >21% oxygen	Total babies in >21% O_2	S
	Necrotising enterocolitis	Necrotising enterocolitis	NS

	Absent	Admission to neonatal intensive or special care	NS
	Absent	Total babies ventilated	NS
	Absent	RDS confirmed by radiography	NS
	Absent	O_2 dependence >28 days	NS
	Absent	O_2 at 36 weeks post conception	NS
	Absent	Abnormal cerebral ultrasonography	NS
	Absent	Composite primary outcome	NS
	Sub-group		
	<32 or ≥32 weeks gestation	<32 or ≥32 weeks gestation	The same pattern of result was found as in the main analysis

NS: Not Significant
S: Significant

Oral vitamin D3, calcium, combination of vitamin D3 and calcium, or placebo for low-trauma fractures in elderly people

Trial identification	Protocol	Reports	Significance level
27	**Primary**		
	All new low-energy fractures or clinical, radiologically supported, vertebral fractures	New low-trauma fractures	NS
	Secondary		
	New radiologically-confirmed fractures	Incidence of all radiologically confirmed fractures	NS
	Death after trial entry	Death	NS
	General health status	Absent	
	Hospital admissions after trail entry	Absent	
	Change of residence category	Absent	
	Falls (within five one-week window periods)	Falls within weeks	NS
	Possible adverse effects	Adverse effects	NS
	New cancer registrations	Absent	
	Deaths attributed to cardiovascular or cerebrovascular disease	Absent	
	New cases of diabetes	Absent	
	Sub-group		**Interaction†**

	Type of fracture (proximal femur, distal forearm, clinical vertebral, other)	Type of fracture (2 groups)	NS
	Age at entry (70-74, 75-79, 80-84, 85 or over)	Age (≥80y or <80y)	NS
	Sex	Sex	NS
	Latitude of recruitment centre (northern vs central vs southern)	Latitude of recruitment centre (2 groups)	NS
	Time of enrolling fracture (within 3 months of recruitment or longer)	Time since fracture (2 groups)	NS
	Dietary calcium (high or low)	Dietary calcium (2 groups)	NS
	Vitamin D exposure (high or low)	Vitamin D intake and exposure (3 groups)	NS
	Weight (<55 kg or not)	Weight (2 groups)	NS
	Level of compliance (completed 2 years or not)	Compliance (>80% or ≤80%)	NS

NS: Not Significant
† The significant level from the interaction test

Intranasal sodium cromoglicate versus intranasal normal saline in children with suspected acute viral upper respiratory tract infection

Trial identification	Protocol	Reports	Significance level
28	**Primary**		
	Suspected acute viral upper respiratory tract infection (SAVURTI) in children measured by the CARIFS composite symptom scale)	Improvement in CARIFS score	NS
	Secondary		
	Use of medication	Absent	
	Side effects	Side effects	NS
	Significant otalgia	Absent	
	Sub-group		
	Age	Age	NS
	Sex	Sex	NS
	History of atopy	History of atopy	NS
	History of infections of the upper respiratory tract	History of infections of the upper respiratory tract	NS
	Duration of symptoms	Duration of symptoms	NS
	Frequency of administration of medication	Absent	
	Clinical features at presentation	Clinical features at presentation	NS

NS: Not Significant

Once versus three-times daily regimens of tobramycin for pulmonary exacerbations of cystic fibrosis

Trial identification	Protocol	Reports	Significance level
29	**Primary**		
	Efficacy:		
	Improvement in forced expiratory volume in 1 second (FEV1)	*Change in forced expiratory volume (FEV1)*	*NS*
	Change in Clinical score	*Clinical score*	*NS*
	C-reactive protein	*C-reactive protein*	*NS*
	Secondary		
	Ototoxicity measured by an audiogram	Deterioration in hearing in audiogram	NS
	Nephrotoxicity by change in creatinine, phosphate, magnesium, and urinary	Changed in creatinine	S
	Sub-group		
	Previous aminoglycoside exposure or not	Absent	
	FEV1 <50% or >50%	Absent	
	Absent	Adults (>16 years old) or children (5-16 years old) (changed in creatinine)	S

NS: Not Significant
S: Significant

Ultrasonographic hip versus clinical assessment in infants with clinical hip instability

Trial identification	Protocol	Reports	Significance level
30	**Primary**		
	Give the baby a hip that is, and remains, functionally unimpaired throughout life	Any hip treatment *	S
	Absent	Treatment with a splint appliance*	S
	Absent	Operative treatment*	NS
	Secondary		
	Amount of treatments required	Absent	
	Independent mobility after the first year of life	Walking by two years	NS
	Sub-group		**Interaction†**
	Initial abnormality (unilateral or bilateral)	Any hip treatment:	
	Diagnosis at entry (dislocated, dislocatable or subluxatable)	*Laterality (2 groups)*	*NS*
		Clinical diagnosis (2 groups)	*NS*
		Treatment with a splint appliance:	
		Laterality (2 groups)	*NS*
		Clinical diagnosis (2 groups)	*NS*
		Operative treatment:	
		Laterality (2 groups)	*NS*

		Clinical diagnosis (2 groups)	*NS*
		Abnormal or borderline appearances on radiograph by 2 years of age:	
		Laterality (2 groups)	*NS*
		Clinical diagnosis (2 groups)	*NS*

* Reported as a primary outcome
NS: Not Significant
S: Significant
† The significant level from the interaction test

Table 3.3 Proportion of trials with discrepancies in the primary outcomes when comparing protocols and published articles (n = 30 trials)

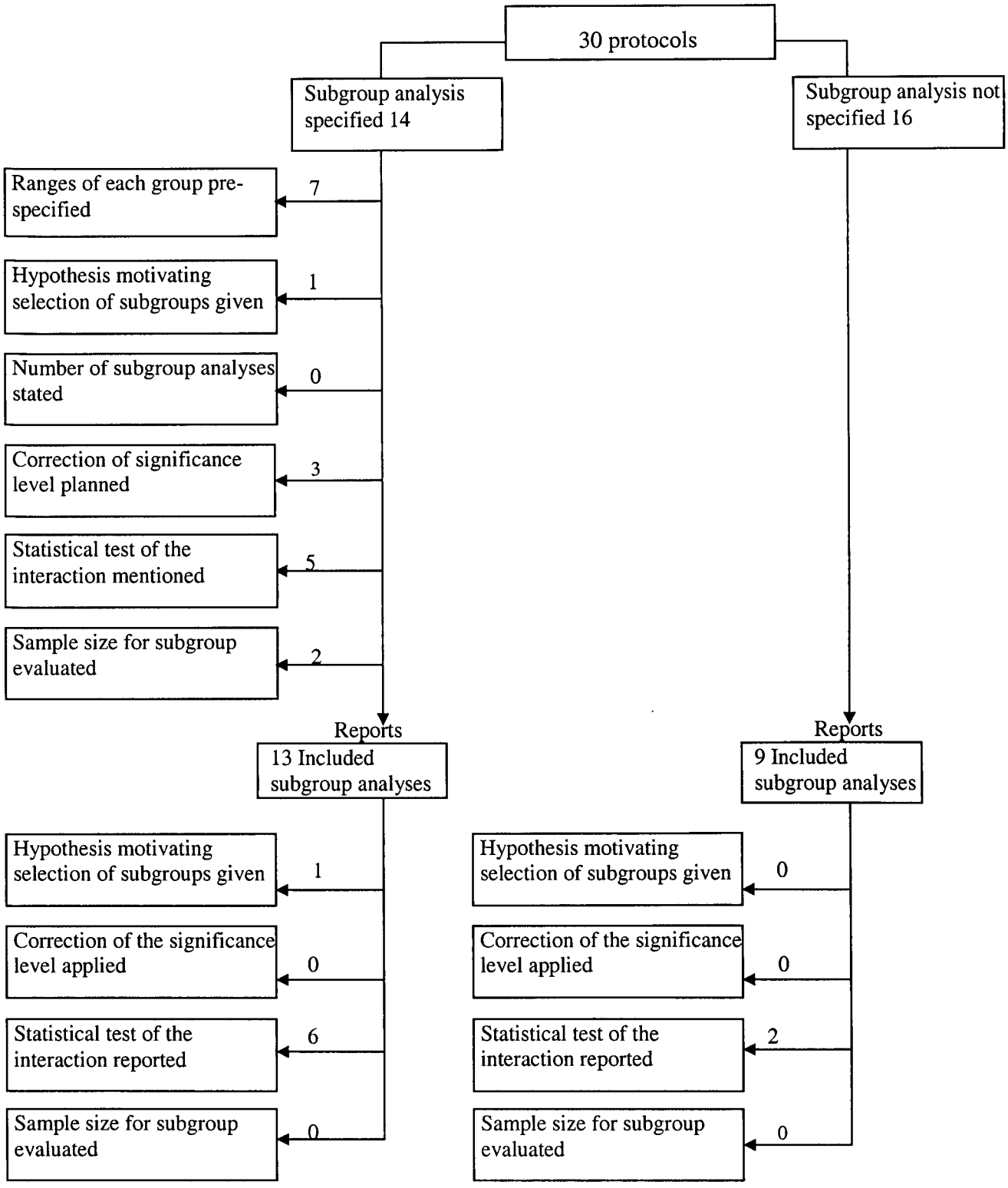
Discrepancies in the primary outcomes in published articles relative to protocols	Trials with discrepancies, No. (%)
Reported but not explicitly defined as primary	21 (70)
Omitted one outcome from published articles *	5 (17)
New primary outcome defined in published articles	7 (23)
Reported as secondary in published articles	2 (7)

*Trials defined >1 primary outcome in protocols and published articles.

Table 3.4 Proportion of trials with discrepancies in the secondary outcomes when comparing protocols and published articles in (n = 30 trials) where the secondary outcomes were defined

Discrepancies in the primary outcomes in published articles relative to protocols	Trials with discrepancies, No. (%)
Omitted 1 outcome from published articles	2 (7)
Omitted 2 outcomes from published articles	11 (37)
Omitted 3 outcomes from published articles	5 (17)
Omitted 5-8 outcomes from published articles	4 (13)
New secondary outcome defined in published articles	11 (37)

Figure 3.1 Specification of sub-group analyses in protocols and report



3.4 Discussion

In this cohort of trials whose protocols had been reviewed and accepted by *the Lancet*, I found evidence of selective reporting of primary and secondary outcomes. There were major discrepancies between the protocols and the reports regarding primary outcomes in one third of trials. There were more primary outcomes in the reports compared to the protocols (60 versus 51) and less secondary outcomes (93 versus 133). The "new" primary outcomes were introduced into the reports with no previous notice in the protocols. There were no statistical adjustments for testing multiple outcomes.

My analyses also show that the pre-specification of sub-groups in the protocols was generally incomplete, with little attention paid to any analysis issues, and only occasional coverage of statistical issues, such as adjustment of significance levels and/or testing for interaction. This deficiency in the protocols was accompanied by frequent deviations between protocols and reports. In more than half of the trials, sub-group analysis was not mentioned in the protocol, but was carried out. In the trials where the sub-groups were pre-specified, there was deviation between protocol and report, with the addition of extra sub-groups being common. There was a lack of reference in the reports to statistical issues, such as adjustment to or discussion of significance levels and testing of interactions.

Although *the Lancet* website provides summaries of all accepted protocols, access to the full protocols required the permission of the investigator, thus introducing the potential for bias. However, I obtained the full protocols for 92% of the accepted protocols and so the potential for bias here was small.

Although I attempted to identify all relevant trial reports, it is possible that some trial reports may have been in preparation or in press and that others simply were not found in the search. For this reason, I may have overestimated the extent of under-reporting. None of the reports gave the reasons why outcomes were omitted or whether they would be presented in later reports, and so it was not possible to tell whether the authors intended to present the outcomes in later publications.

The number of trials in the study was small, and so the estimates of the frequency of various features are approximate. In addition, the study cohort comprised protocols accepted by *The Lancet* and was not a random sample of all trial protocols. For this reason, the extent to which these results can be generalised to other clinical trials is open to question. However, I would expect that protocols accepted by *the Lancet* would represent those of higher scientific quality and that the problem of selective reporting and inadequate specification in the protocols would be worse in other trials.

Concerning multiple outcomes, the conclusions in this review are similar to those reported by Chan *et al* (3, 4). However, previous reviews did not investigate the reporting of sub-group analyses.

Chan *et al* (3, 4) classified the level of outcome reporting in four groups (full, incomplete, qualitative or unreported) based on the published reports of each trial. This classification was set up to assess the suitability for use in meta-analysis. In this study, the Delphi survey focused on selective reporting as such, and at this stage fitness for meta-analysis was not in question.

This review has examined the protocols and/or reports of trials which were peer reviewed by statisticians and accepted by *the Lancet*. However, the issue of precise definition of the primary and secondary outcomes and sub-groups in the protocols does not appear to have been sufficiently addressed. Deviations in the published reports were found, and no report mentioned the discrepancies or the reasons, although it had been recommended to describe deviations from protocols in the published reports (3, 69). The most common reasons given by trialists for not publishing all outcomes were lack of statistical significance, journal space restrictions and lack of clinical importance (70).

Many authors (71) argue that any sub-group analyses conducted should be restricted to those proposed before data collection on the basis of known biological mechanisms or in response to findings from previous studies. This is to guard against the potential for post hoc emphasis on the 'most interesting' across many sub-group analyses (72). However, if a sub-group analysis was not originally planned, but was decided on during the life of the study in response to new results from other studies, it makes good scientific sense to examine this, irrespective of what was stated in the protocol (71). Seldom was there a biological rationale offered for the sub-group analyses, either in the protocols or in the published reports.

Correction for multiple testing (sub-group, multiple outcome) was very seldom identified, but failure to adjust the significance level can lead to a serious problem. A correction for multiple testing was mentioned in just three of the trial protocols, none of them reporting any applied correction. Bonferroni adjustment of the significance level is one option for such adjustment, but it assumes all groups are mutually exclusive, i.e. patients must be included in one category only. This review shows that sub-group analyses are not always mutually exclusive, so appropriate correction is difficult.

The use of formal tests of interaction, which directly examine the difference between intervention effects in different sub-groups, can reduce this problem. This approach involves one statistical test irrespective of the number of sub-groups. The single interaction test may partially overcome the concerns of a false positive conclusion of a treatment, but will be underpowered if the sample size is not set for the interaction test (8). In this study, the interaction test was carried out in just 8 out of 22 (36%) trials where the results of sub-groups were mentioned in the reports.

The potential problem from unscheduled or inappropriate sub-group analyses

The results demonstrate that in nine of the sixteen (56%) trials, sub-group analyses were reported but had not been mentioned in the trial protocols. Neither the statistical power nor the significance level had been adjusted to detect sub-group treatment effects. Fortunately, none of these trials recommended using the treatment on the basis of the sub-group analysis, but there is no reason to think this will not happen in the future. It is important to note that this is a problem not only in non pre-specified sub-group analysis but also in pre-specified analyses carried out inadequately.

In terms of the potential for misleading inferences, there are three possible scenarios. Firstly, the overall and the sub-group results have the same direction and similar magnitude of effect, either significant or non-significant. Thus, the conclusion will be the same and there is little risk of misleading inferences. Secondly, the overall result is significant, whereas the sub-group result is non-significant. This almost inevitably arises because of the small sample size in the sub-group leading to reduced statistical power to detect an effect of the intervention. This may lead to a group of patients being denied an effective treatment (a false negative conclusion) if "non-significant" is taken as "no effect". Thirdly, the overall

result is not significant and one or more sub-group analyses are significant. The significant result could be by chance, especially if no interaction test is applied. This is potentially the worst case, since it may lead to a group of patients being given an ineffective treatment (a false positive conclusion), particularly if the emphasis and recommendations were on the sub-group results. For instance, my study included a multi-centre randomised controlled trial comparing once daily and three times daily dosing regimens in groups of patients with cystic fibrosis and a chronic *Pseudomonas aeruginosa* infection (64).

It was mentioned in the protocol of this trial that the primary endpoint would be efficacy, as measured by improvement in forced expiratory volume (FEV_1). Clinical efficacy would also be measured by changes in clinical score and C reactive protein with treatment. The study was powered to detect equivalent efficacy. The secondary endpoints would be measures of ototoxicity and nephrotoxicity. Ototoxicity would be measured by an audiogram and nephrotoxicity by changes in creatinine, phosphate, magnesium, and urinary. Sub-group analysis was planned by previous aminoglycoside exposure and by comparing those patients with FEV1 <50% predicted with those who have an FEV1 of 50% predicted or above at enrolment.

The publish article reported the findings of the mean change in FEV_1 (% predicted) over the 14 days of treatment which was similar on the two regimens. There was no significant difference in % change in creatinine from baseline. However, the article reported the findings for young children separately, for whom there was evidence of a "significant" treatment effect. The authors pointed out that "in children, once daily treatment was significantly less nephrotoxic than thrice daily (mean % change in creatine -4.5% [once daily] *vs* 3.7% [thrice daily]; adjusted mean difference -8.0%, 95 CI -15.8 to -0.4)". They concluded that "the once daily regimen might be less nephrotoxic in children". Here a

conclusion was drawn from a sub-group analysis that was not originally planned in the protocol. It was measured for a secondary outcome and no statistical interaction test was done to indicate sufficient evidence that the intervention's effect was different in the sub-group. Therefore, this result could be due to chance, and further investigation would be needed.

When the interaction test is significant, how should the sub-group effects be interpreted? An example from this review is given. A randomised controlled trial tested the efficacy of full treatment with glasses and patching versus glasses only versus no treatment in preschool children with unilateral visual acuity defects of 6/9 to 6/36 (58).

The protocol of this study stated the primary outcome as assessment of uncorrected LogMAR visual acuity. Sub-group analyses were planned to test for heterogeneity between centres and at different initial acuities.

The results showed that children in the full and glasses treatment groups had better visual acuity at follow-up than children who received no treatment, but the overall treatment effect was small. The study report gave much space to sub-group analyses for children with moderate and mild acuity at recruitment, and gave the results of a statistical test of interaction. The authors concluded that "The effects of treatment depended on initial acuity: full treatment showed a substantial effect in the moderate acuity group (6/36 to 6/18 at recruitment) and no significant effect in the mild acuity group (6/9 to 6/12 at recruitment) ($p= 0.006$ for the linear regression interaction term)".

Although the finding of the interaction test was reported and was significant; the range of values defining moderate or mild acuity at baseline was not specified in the protocol of this

trial. Thus, how one can be sure that this sub-group analysis was the only one and not one of several exploratory sub-group analyses that were conducted but not reported. There is a real risk of an exaggerated false positive or an over-estimated treatment effect. Such sub-group findings should have been a basis for further research from similar trials rather than the basis for change in national policy.

Some proposed issues to control inappropriate sub-group analyses and selective reporting in clinical trials discussed in Chapter 6.

References

1 Hutton J, Williamson P. Bias in meta-analysis due to outcome variable selection within studies. *Applied statistics* 2000; 49:359-370.

2 Hahn S, Williamson P, Hutton J, Garner P, Flynn EV. Assessing potential for bias in meta-analysis due to selective reporting of subgroup analysis within studies. *Statistics in Medicine* 2000; 19:3325-3336.

3 Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004; 291:2457-65.

4 Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomised trials funded by the Canadian Institutes of Health Research. *CMAJ* 2004; 17(7):735-740.

5 Gelber RD, Goldhirsch A. Inerpretation of results from subset analyse within overviews of randomised clinical trials. *Statistics in Medicine* 1987; 6:371-378.

6 Tannock IF. False-positive results in clinical trials: multiple significance tests and the problem of unreported comparisons. *Journal of the National Cancer Institute* 1996; 88:206-207.

7 Chalmers I. Unrerreporting research is scientific misconduct. *JAMA*. 1990; 263:1405-14080.

8 Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials: a survey of three medical journals. *The New England Journal of Medicine* 1987; 317:426-432.

9 Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. *Health Technology Assessment* 2000; 4:1-115.

10 Hahn S, Williamson P, Hutton J. Investigation of within-study selective reporting in clinical research: follow-up of applications submitted to local research ethics committee. *Journal of Evaluation in Clinical Practice* 2002; 8:353-359.

11 Pocock SJ. Clinical Trials: A Practical Approach; John Wiley & Sons: Chichester, 1983.

12 Sleight P. Commentary debate: subgroup analyses in clinical trials – fun to look at, but don't believe them! *Current controlled trials in cardiovascular medicine* 2000; 1:25-27.

13 Brookes ST, Whitley E, Peters TJ, et al. Subgroup analyses in randomised controlled trials: quantifying the risks of false positives and false negatives. *Health Technology Assessment* 2001; 5:1-56.

14 Lu M, Lyden PD, Brott TG, Hamilton S, Broderick JP, and Grotta JC. Beyond subgroup analysis: improving the clinical interpretation of treatment effects in stroke research. *Journal of Neuroscience Methods* 2005; 143:209-216.

15 Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: Some lessons from major cardiovascular trials. *American Heart Journal* 2000; 139:952-61.

16 Begg C, Cho M, Eastwood S et al. improving the quality of reporting of randomized controlled trials. Journal of the American Medical Association 1996; 276,637-639.

17 European Agency for the Evaluation of Medical Products ICH Guidelines. Human Medicine Evaluation Unit, ICH-Technical Coordination, London 1998.

18 Staessen. JA, Bianchi. G. Registration of trials and protocols. *Lancet* 2003; 362:1009-10.

19 Krleža-Jerić K, Chan AW, Dickersin K, Sim I, Grimshaw J, Gluud C. Principles for international registration of protocol information and results from human trials of health related interventions: Ottawa statement (part 1). *BMJ* 2005;330:956-958.

20 Abbasi K. Compulsory registration of clinical trials. *BMJ* 2004; 329:637-638.

21 Godlee F. Publishing study protocols: making them more visible will improve registration, reporting and recruitment. *BMC News Views* 2001; 2:4.

22 Lassere M, Johnson K. The power of the protocol. *Lancet* 2002; 360:1620-1622.

23 Hawkey CJ. Journal should see original protocols for clinical trials. *BMJ* 2001; 323:1309.

24 Jones G, Abbasi K. Trial protocols at the BMJ. *BMJ* 2004; 329:1360.

25 De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *Lancet* 2004; 364: 911-2.

26 http://www.thelancet.com/journals/lancet/misc/protocol/protocolreviews

27 Holmberg L, Anderson H. HABITS (hormonal replacement therapy after breast cancer—is it safe?), a randomised comparison: trial stopped. *Lancet* 2004; 363:453-455.

28 Guillou PJ, Quirke P, Thorpe H, Walker J, Jayne DG, Smith AM, et al. Short-term endpoints of conventional versus laparoscopic-assisted surgery in patients with colorectal cancer (MRC CLASICC trial): multi-centre, randomised controlled trial. *Lancet* 2005; 365:1718-1726.

29 O'Brien J, Duncan H, Kirsh G, Allen V, King P, Hargraves R, et al. Prevention of pulmonary embolism and deep vein thrombosis with low dose aspirin: Pulmonary Embolism Prevention (PEP) trial. *Lancet* 2000; 355:1295-1302.

30 Progress Collaborative Group. Randomised trial of a perindopril-based blood-pressure-lowering regimen among 6105 individuals with previous stroke or transient ischaemic attack. *Lancet* 2001; 358:1033-1041.

31 Hasegawa Y, Yamaguchi T, Omae T, Woodward M, Chalmers J. Effects of perindopril-based blood pressure lowering and of patient characteristics on the

progression of silent brain infarct: the perindopril protection against recurrent stroke study (PROGRESS). *Hypertens Res* 2004; 27:147-156.

32 Chapman N, Huxley R, Anderson C, Bousser MG, Chalmers J, Colman S, Davis S, Donnan G, MacMahon S, Neal B, Warlow C and Woodward M. Effects of perindopril-based blood pressure lowering regimen on the risk of recurrent stroke according to stroke subtype and medical history: the Progress Trial. *Stroke* 2004; 35:116-21.

33 Marteau T, Senior V, Humphries S, Bobrow M, Cranston T, Crook M et al. Psychological Impact of Genetic Testing for Familial Hypercholesterolemia Within a Previously Aware Population: A Randomized Controlled Trial. *American Journal of Medical Genetics* 2004; 128A:285–293.

34 Rigg JR, Jamrozik K, Myles PS, Silbert BS, Peyton PJ, Parsons RW, et al. Epidural anaesthesia and analgesia and outcome of major surgery: a randomised trial. *Lancet* 2002; 359:1276-1282.

35 International Subarachnoid Aneurysm Trial (ISAT) Collaborative Group. International Subarachnoid Aneurysm Trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: a randomised trial. *Lancet* 2002; 360:1267-1274.

36 International Subarachnoid Aneurysm Trial (ISAT) Collaborative Group. International Subarachnoid Aneurysm Trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: a randomised comparison of effects on survival, dependency, seizures, rebleeding, subgroups, and aneurysm occlusion. *Lancet* 2005; 366:809-817.

37 Zajicek J, Fox P, Sanders H, Wright D, Vickery J, Nunn A, et al. Cannabinoids for treatment of spasticity and other symptoms related to multiple sclerosis (CAMS study): multi-centre randomised placebo-controlled trial. *Lancet* 2003; 362:1517-1526.

38 Hobart JC, Riazi A, Thompson AJ, Styles IM, Ingram W, Vickery PJ, et al. Getting the measure of spasticity in multiple sclerosis: the multiple sclerosis spasticity scale (MSSS-88). *Brain* 2006; 129:224-234.

39 Kluin-Nelemans H, Buck G, Cessie S, Richards S, Beverloo HB, Falkenburg J, et al. Randomized comparison of low-dose versus high-dose interferon-alfa in chronic myeloid leukemia: prospective collaboration of 3 joint trials by the MRC and HOVON groups. *Blood* 2004; 103:4408-4415.

40 Cuzick J, Szarewski A, Cubie H, Hulman G, Kitchener H, Luesley D, et al. Management of women who test positive for high-risk types of human papillomavirus: the HART study. *Lancet* 2003; 362:1871-1876.

41 Mendelow AD, Gregson BA, Fernandes HM, Murray GD, Teasdale GM, Hope DT, et al. Early surgery versus initial conservative treatment in patients with spontaneous supratentorial intracerebral haematomas in the International Surgical Trial in Intracerebral Haemorrhage (STICH): a randomised trial. *Lancet* 2005; 365:387-397.

42 O'Grady J, Burroughs A, Hardy P, Elbourne D, Truesdale A. Tacrolimus versus microemulsified ciclosporin in liver transplantation: the TMC randomised controlled trial. *Lancet* 2002; 360:1119-1125.

43 Fallowfield L, Jenkins V, Farewell V, Saul J, Duffy A, Eves R. Efficacy of a Cancer Research UK communication skills training model for oncologists: a randomised controlled trial. *Lancet* 2002; 359:650-656.

44 The GRIT study group. Infant wellbeing at 2 years of age in the Growth Restriction Intervention Trial (GRIT): multi-centred randomised controlled trial. *Lancet* 2004; 364:513-520.

45 Bueving H, Bernsen R, Jongste J, Suijlekom-Smit L, Rimmelzwaan G, Osterhaus A, et al. Influenza vaccination in children with asthma. *Am J Respir Crit Care Med* 2004; 169:488-493.

46 Bueving H, Bernsen R, Jongste J, Suijlekom-Smit L, Rimmelzwaan G, Osterhaus A, et al. Does influenza vaccination exacerbate asthma in children? *Vaccine* 2004; 23:91-96.

47 Halmerbauer G, Gartner C, Schierl M, Arshad H, Dean T, Koller DY, et al. Study on the Prevention of Allergy in Children in Europe (SPACE): Allergic sensitization in children at 1 year of age in a controlled trial of allergen avoidance from birth. *Pediatr Allergy Immunol* 2002; 13 (Suppl. 15):47–54.

48 Gülmezoglu AM, Villar J, Ngoc NTN, Piaggio G, Carroli G, Adetoro L,et al. WHO multi-centre randomised trial of misoprostol in the management of the third stage of labour. *Lancet* 2001; 358:689-695.

49 Stephenson J, Strange V, Forrest S, Oakley A, Copas A, Allen E, Babiker A, Black S, Ali M, Monteiro H, Johnson A. Pupil-led sex education in England (RIPPLE study): cluster-randomised intervention trial. *Lancet* 2004; 364:338-346.

50 Field D, Elbourne D, Truesdale A, Grieve R, Hardy P, Fenton A, et al. Neonatal Ventilation With Inhaled Nitric Oxide Versus Ventilatory Support Without Inhaled Nitric Oxide for Preterm Infants With Severe Respiratory Failure: The INNOVO Multi-centre Randomised Controlled Trial. *Pediatrics* 2005; 115:926-936.

51 Ahluwali J, Tooley J, Cheema I, Sweet DG, Curley AE, Halliday HL, et al. A dose response study of inhaled nitric oxide in hypoxic respiratory failure in preterm infants. *Early Human Development* 2006; 82: 477-483.

52 CAVATAS investigators. Endovascular versus surgical treatment in patients with carotid stenosis in the Carotid and Vertebral Artery Transluminal Angioplasty Study (CAVATAS): a randomised trial. *Lancet* 2001; 357:1729-1737.

53 McCabe D, Pereira AC, Clifton A, Bland JM, Brown MM on behalf of the CAVATAS investigators. Restenosis after carotid angioplasty, stenting or endarterectomy in the Carotid and Vertebral Artery Transluminal Angioplasty Study (CAVATAS). *Stroke* 2005; 36:281-286.

54 The Magpie Trial Collaborative Group. Do women with pre-eclampsia, and their babies, benefit from magnesium sulphate? The Magpie Trial: a randomised placebo-controlled trial. *Lancet* 2002; 359:1877-1890.

55 CRASH trial collaborators. Effect of intravenous corticosteroids on death within 14 days in 10 008 adults with clinically significant head injury (MRC CRASH trial): randomised placebo-controlled trial. *Lancet* 2004; 364:1321-1328.

56 CRASH trial collaborators. Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury—outcomes at 6 months- *Lancet* 2005; 365:1957-1959.

57 Oommen A, Lambert PC, Grigg J. Efficacy of a short course of parent-initiated oral prednisolone for viral wheeze in children aged 1–5 years: randomised controlled trial. *Lancet* 2003; 362:1433-1438.

58 Clarke MP, Wright CM, Hrisos S, Anderson JD, Henderson J, Richardson S R. Randomised controlled trial of treatment of unilateral visual impairment detected at preschool vision screening. *BMJ* 2003;327:1251-1255.

59 Barwell JR, Davies CE, Deacon J, Harvey K, Minor J, Sassano A, et al. Comparison of surgery and compression with compression alone in chronic venous ulceration (ESCHAR study): randomised controlled trial. *Lancet* 2004; 363:1854-1859.

60 Kenyon S, Taylor D, Tarnow-Mordi W. Broad-spectrum antibiotics for preterm, prelabour rupture of fetal membranes: the ORACLE I randomised trial. *Lancet* 2001; 357:979-988.

61 Kenyon S, Taylor D, Tarnow-Mordi W. Broad-spectrum antibiotics for spontaneous preterm labour: the ORACLE II randomised trial. *Lancet* 2001; 357:989-994.

62 The RECORD Trial Group, Oral vitamin D3 and calcium for secondary prevention of low-trauma fractures in elderly people (Randomised Evaluation of Calcium Or vitamin D, RECORD): a randomised placebo-controlled trial. *Lancet* 2005; 365:1621-1628.

63 Butler CC, Robling M, Prout H, Hood K, Kinnersley P. Management of suspected acute viral upper respiratory tract infection in children with intranasal sodium cromoglicate: a randomised controlled trial. *Lancet* 2002; 359:2153-2158.

64 Smyth A, Tan KHV, Hyman-Taylor P, Mulheran M, Lewis S, Stableforth D, et al. Once versus three-times daily regimens of tobramycin treatment for pulmonary exacerbations of cystic fibrosis-the TOPIC study: a randomised controlled trial. *Lancet* 2005; 365:573-578.

65 Elbourne D, Dezateux C, Arthur R, Clarke N, Gray A, King A, et al. Ultrasonography in the diagnosis and management of developmental hip dysplasia (UK Hip Trial): clinical and economic results of a multi-centre randomised controlled trial. *Lancet* 2002; 360:2009-2017.

66 White I, Elbourne D. Assessing subgroup effects with binary data: can the use of different effect measures lead to different conclusion? *BMC Medical Research Methodology* 2005; 5:15-20.

67 Gardner F, Dezateux C, Elbourne D, Gray A, King A, Quinn A. The hip trial: psychosocial consequences for mothers of using ultrasound to mange infants with development hip dysplasia. *Arch Dis Child Fetal Neonatal Ed* 2005; 90:F17-F24.

68 Gray A, Elbourne D, Dezateux C, King A, Quinn A, Gray A. Economic evaluation of ultrasonography in the diagnosis and management of developmental hip dysplasia in the United Kingdom and Ireland. *J Bone joint Surg Am* 2005; 87:2472-2479.

69 Goldbeck-Wood S. Changes between protocol and manuscript should be declared at submission. *BMJ* 2001;322:1460-1.

70 Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ* 2005;330:753-6.

71 Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. *Medical Journal of Australia* 2004; 180(6):289-291.

72 Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trials reporting: current practice and problems. *Statistics in Medicine* 2002; 21:2917-2930.

CHAPTER 4

Statistical techniques to detect data fabrication and falsification: systematic review

4.1 Introduction

Data fabrication (making up data values) and falsification (changing data values) might be classified as one of the most severe types of scientific misconduct. Conclusions drawn from any research where such misconduct has taken place may be unreliable. This is serious misconduct to the point that Buyse *et al* (1) used the term "fraud" specifically to refer to data fabrication and falsification.

It has been shown from the Delphi survey (chapter 2) that data fabrication and data falsification were judged by the experts to be a relatively uncommon problem. Its actual prevalence is unknown. However, they also considered it to have considerable potential impact on the outcome of clinical trials if it does occur (2). Thus, the detection of such misconduct is an important issue.

Monitoring visits to the clinical centres participating in a trial is one approach for fraud detection (3,4). In some circumstances, such monitoring should be routine and has found instances of fraud during these visits (5). However, it is expensive and difficult to verify all items especially if the volume of data being checked is very high. Where there are particular grounds to suspect misconduct, it is of help to submit clinical trial data to more extensive checks. Statistical techniques for fraud detection can be used as screening mechanisms or for further investigation of data that fall under suspicion. They could even be implemented more

easily than the monitoring approach especially with modern computer programs for statistical data analysis, so long as the primary data can be obtained.

To identify statistical techniques that could be used for the detection of fraudulent data, a systematic review was performed. It is believed that this review is an aid to conceptualising which techniques have potential for use, as well as providing a guide to researchers of the context within which each could be used in detection.

4.2 Methods

4.2.1 Inclusion criteria

All English language reports presenting statistical techniques for the detection of fraudulent data were sought. Eligible studies were not restricted to clinical trial data, and included any study in any population.

4.2.2 Search strategy

The following electronic databases were searched: EMBASE, Web of KNOWLEDGE, and PubMed.

The following search terms were used for EMBASE from 1980 to August 2004

1. Fraud$
2. Scientific misconduct
3. Falsify$.
4. Statistic$ OR *STATISTICS
5. (1 OR 2 OR 3) AND 4

The following search terms were used for Web of SCIENCE from 1981 to August 2004

1. Fraud* OR scientific misconduct OR falsif*

2. Statistic*
3. #1 AND #2

The following search terms were used for PubMed from 1980 to August 2004

1. ("Fraud/prevention and control"[MeSH] OR "Fraud/statistics and numerical data"[MeSH]) AND "Fraud"[MeSH] OR "Scientific Misconduct/statistics and numerical data"[MeSH] OR "Scientific Misconduct"[MeSH] Field: MeSH Major Topic
2. "Statistics"[MeSH] Field: MeSH Major Topic
3. #1 AND #2 Field: MeSH Major Topic

The reference lists of relevant papers were searched for other possible pertinent papers.

4.2.3 Identification of records and data extraction

Each record identified by the search strategy was screened for eligibility by examining titles, abstracts, and keywords. The full text of all potentially relevant reports was then obtained for further assessment for inclusion. Data were extracted from eligible reports using a data abstraction form developed specifically for this review.

4.3 Results

The search identified 316 potentially relevant articles. Of these, 16 articles were found to meet the inclusion criteria, involving 19 statistical techniques for the detection of fraud. The other 300 papers did not include any statistical techniques.

Statistical techniques that can be used for fraud detection fall into two primary classes: analytical techniques and graphics that supplement them. The analytical techniques can produce a measure or a test result. As well as formal statistical tests under null hypotheses performed for confirmation of fraud, preliminary analyses or summary measures can be presented to show the properties of the data under investigation. Some references included

looking for features, patterns, or trends in the data that would be unlikely to occur in genuine data.

4.3.1 Statistical properties of fraudulent data

The statistical examination of data suspected for fraud can be focused on the data set as a whole or on sub-sets of the data. It is helpful to look at descriptive statistics such as the prevalance of a binary variable or the mean, the median, the standard deviation, and both the minimum and maximum values of a continous variable to reveal discrepancies if they are compared with another dataset or previous studies. Univariate values of such variables tend to be fabricated with a series that fall close to the mean. Consequently, fraudulent data often have a smaller variance than is seen with real data (6,7). Outliers (values unusually far from the overall mean) tend to be removed rather than inserted (7) or outliers may occur frequently or be clustered in one centre (6). The range and the kurtosis of the distribution may be helpful in detecting outliers. The kurtosis is able to detect departures from a normal distribution e.g. a uniform distribution will show a high value of kurtosis (7).

Many biological variables are expected to be distributed fairly normally or log normally. The distribution of fraudulent data with values being invented tends to be relatively flat (7). The distribution of invented data tends to be normal but with round numbers for the mean and standard deviation (8).

A reduction in the variance of observations over time gives a suspicion of data fabrication. For example, a multi-centre study was conducted using an animal model of myocardial infarction (9). The data showed that the heart rates of consecutively treated animals had far too little variability. Insufficient variability over time was discussed by Buyse et al (1) who

quoted an example from Scherrer (10) and showed that any such reduction may reveal a problem in the data.

Another way of checking fraudulent data follows from the fact that people have difficulty in generating long sequences of random digits (11). If measurements are recorded with reasonable precision, the observed counts of each final digit (0-9) should be random and follow a uniform or rectangular, distribution. The phenomenon of "digit preference" is well known when human beings make observations that are not perfectly precise. This happens because in rounding the last digit, certain values are preferred such as zero and five. This typically happens with blood pressure and is not necessarily indicative of any fraud.

Deviation of the final digit from the uniform distribution, as would be expected, may suggest that something is amiss (12), e.g. some people will chose 4 and 7 preferentially. For some measures such as blood pressure, the preferential use of 0 and 5 as end digits is commonly seen in genuine data (13). Preece (12) describes a number of possible sources of terminal digit preference other than misconduct.

The use of the "stem and leaf" plot is helpful for examining digit preference (6,13,14). A stem and leaf plot is like a histogram on its side, with the "stem" being the most significant digits, and the "leaves" being the least significant digits. Because it retains all the data, unlike the histogram, which groups the data, the last digits can be seen, and instances of digit preference can be seen clearly. If digit preference is suspected, then a histogram of the final digit or a separate one of the penultimate digit can be helpful.

Benford's law (1,15,16) may be used to check the randomness of the first digit of all real numbers reported by a single individual (or a single centre). Benford's law states that in

listing, tables of statistics, etc., the probability that the first digit will be a "1" is about 0.301, rather than 0.11 as might be expected if all digits were equally likely. In general, the law says that the probability of the first significant digit being a "D" is

$$p(D) = \log_{10}\left(1+\frac{1}{D}\right)$$

This implies that a number in the data is more likely to begin with a smaller digit than a larger digit.

This law requires that the data have a wide range, more than 3 orders of magnitude. It does not apply to for example systolic blood pressures where the first digit has an entirely different distribution. It has proved more useful in the detection of financial fraud and has not yet been shown to be useful in medical research.

Plotting means against the variances for the variables of interest for different centres or for different investigators may indicate that one of a set has a different pattern of results (14).

Autocorrelation (1) can be used to show the correlation with the immediately prior observation, then the observation two before, 3 before etc. These correlation coefficients can be plotted to see the relation between observations as a function of sequence of entry to the trial. If the data do not hold the property of independence and there is much dependence between successive observations, the autocorrelations will be large. This shows problems because human beings cannot invent truly random number.

Some fraud only becomes apparent when two variables are compared. It is more difficult for someone generating fraudulent data to retain the nature of real data when viewed in two dimensions and the relationships between variables tend to disappear (7). In the animal

study, (9) the relationship between ventricle weight and dog weight was not as close in one lab as in the other one and the relationship between infarct size and collateral blood flow to the heart was even worse.

Alternatively, correlation coefficients in fabricated data sets can be greater than that found in real data sets. Danesh and Kooshkghazi (17) aimed to study the characteristics of real and fabricated data sets in term of the association between two variables. Two examples are presented on two different settings: first, when there is high correlation coefficient between variables (weight and height), second, when the variables are not correlated. (The authors' example of the latter is birth weight and gestational age, which in fact are naturally correlated). The outcomes from fabricated data sets were compared with the results from two real data sets and with appropriate simulated data sets. The correlation coefficient in the fabricated data was always higher than in the real one. The author wrote, "The results indicate that high correlation coefficients can be considered as a potential sign of data fabrication." This of course depends on good knowledge of what correlations occur in real data.

One can check if a regression, which obtains as good a fit of a linear model to the data as possible, makes biological sense. Finding association between variables as significant, which are not thought to be biologically associated, may indicate problems with the data (7).

Plotting data values against time or the order in which they are entered on record forms, can show trends in the data, which are consistent with non-random components. Bailey (9) shows some graphs where the data collected by an individual extended over a time period that included both genuine and invented data, and where the time periods when fraud occurred stood out clearly.

Chernoff faces are a method of visualizing multidimensional data (18). While Chernoff faces provide an effective way of revealing rather complex relations not always visible from simple correlations, it would not be feasible to produce such plots for each patient in large studies. Star plots are another way of displaying multivariate data. They can accommodate as many variables as necessary with equal emphasis. They illustrate the data by representing each variable by a radial line of length proportional to the size of response, the ends of these lines are joined, producing a star (13). Chernoff faces and star plots can give useful indications of variation if used to plot mean responses for each centre and can reveal unexpected patterns in the data.

It is also useful to look for the change in digit preference over time. A cumulative sum plot can reveal if any change of investigator or invention of patient's readings after a certain time point. The deviation from the straight line indicates the frequency is changing (13).

The examination of residuals is useful to look at influential or outlying observation. Plots of residuals against subject sequance number may be checked. Bailey (9) showed such graphs were particularly revealing.

The Mahalanobis distance is a measure of the distance of an observation of several variables from a multivariate mean. The Mahalanobis distance is computed by standardizing the variables of interest (subtracting the mean and dividing by the standard deviation), and summing the squares of these standardized quantities for each individual. This distance follows a *chi*-square distribution, approximately. It can detect multi-dimensional inliers (values falling close to the multivariate mean) or multi-dimensional outliers (values falling far from the multivariate mean). Evans (7) explains the use of the Mahalanobis distance for a set of data to which two inliers have been added.

Cluster analysis can be applied where there is a possibility that the results have been obtained, for example, by splitting specimens from a single patient to generate several samples supposedly from different patients. This can be done for several patients to obtain larger numbers of patients (1,6). Cluster analysis can then show that these apparently different patients are too similar to one another. This can only demonstrate misconduct when genuine duplicate observations are also available.

Discriminant analysis and can be used to explore the data and possibly detect outlying sites (6) specifically, it would not be expecting observations to differ between sites other than through the play of chance. Although these authors mentioned this technique, they did not give details about its use for detection or confirmation of fraud. Table 4.1 summaries these techniques.

Table 4.1 The nature of fraudulent data and the technique to detect it

Nature of fraudulent data	How to detect
Outliers	The range
Terminal digit preference	Stem and leaf Histogram
First digit preference	Benford's law
Shape of the data	Histogram kurtosis
Change in digit preference over time	Cumulative sum plot
Dependence between successive observations	Autocorrelation
Different pattern between groups	Plotting means vs variances Cross tabulation
Inconsistent relationships	Correlation Regression Cross tabulation Scatter plot Residual plot
Trends in the data (non-random)	Plotting data values vs time
Influential observation	Residual plot
Multivariate (inliers or outliers)	Mahalanobis distance Discriminant analysis Cluster analysis
Unexpected pattern in multivariate data	Chernoff faces plot Star plot
Duplicated data	Cluster analysis

4.3.2 Statistical tests to indicate fraud

The final digit (0-9) of any measurement with several significant digits tends to have a uniform distribution. If digit preference is suspected, a chi square test can be used to examine the uniformity (7,13). This test is also helpful in a comparison between genuine and suspect data in terms of similarity of distribution. Examining the pattern of digit preference by investigator in a multi-centre trial or by randomised treatment group can show differences that, at the least, require further investigation.

Comparing the means of two groups using a t test in randomised trials at baseline is not useful when a trial is truly randomised since any difference is certainly due to chance. However, when fraud is suspected then use of t-tests or other comparisons can be indicative of a problem (19). Statistically significant results using a t-test may be found when data have been fabricated or falsified after randomisation (7).

The runs test can be used to decide if successive observations arise from a random process and the consecutive values are independent of one another (1). A run is defined as a series of increasing values or a series of decreasing values, or the number of successive values above or below the median, and that is the length of the run. In a random data set, the probability that the (N+1) th value is larger or smaller than the Nth value follows a binomial distribution which forms the basis of the runs test.

4.4 Conclusion

Randomised controlled clinical trials are among the strongest designs of medical research for making causal inference, and any misconduct undermines their validity.

Since statistical techniques can be applied to fraudulent data, biostatisticians should be involved in searching for such fraud. It has been suggested that exact statistical details of how they are detected should not be widely published to avoid a rise in the complexity of fraud (7), or giving those wishing to pervert science an opportunity to learn how to avoid detection.

Many of the techniques described in this chapter identified general departures from genuine data. Although, in some cases it is important to state that there can be a variety of explanations other than fraud.

In the context of clinical trials, usually involving several centres, fabricated data from particular centres are easier to detect. The unusual values or patterns in the data might be compared across centres, investigators, treatment groups, or with previous studies. This can be done in term of measures such as, variability, digit preference, histograms or the relationship of various pairs of variables by the use of correlation, regression, scatter plot, or cross tabulation (14).

Sometimes inspecting appropriate graphs could be more informative than applying statistical techniques and tests (7,13).

Reference:

1 Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B, et al. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Statistics in Medicine* 1999;18:3435-51.

2 Al-Marzouki S, Roberts I, Marshall T, Evans S. The effect of scientific misconduct on the results of clinical trials: A Delphi survey. *Contemporary Clinical Trials* 2005;26:331-337.

3 Mackintosh DR, and Zepp VJ. Detection of negligence, fraud, and other bad faith efforts during field auditing of clinical trial sites. *Drug Information Journal* 1996;30:645-653.

4 Schwarz RP. Maintaining integrity and credibility in industry-sponsored clinical research. *Controlled Clinical trials* 1991;12(6):753-760.

5 Seachrist L. NIH trial monitoring: hit or miss? *Science* 1994;264:1534-1537.

6 Collins M, Evans S, Moynihan J, Piper D, Thomas P, Wells F. Statistical techniques for the investigation of fraud in clinical research; Report of the ABPI Fraud Statistics Working Party, February 1993.

7 Evans S. Statistical aspects of the detection of fraud. In: Lock S, Wells F, Farthing M, eds. *Fraud and misconduct in medical research.* 3rd ed. London: *BMJ* Publishing Group, 2001:186-204.

8 Dawson RJ, Mac G. How many light bulbs does it take to generate a data set?, *American Statistician* 1996;50:247-249.

9 Bailey KR. Detecting fabrication of data in a multicentre collaborative animal study *Controlled Clinical Trials* 1991;12:741-752.

10 Scherrer B. L'apport de la biometrie, in La Fraude dans les Essais Cliniques, *Medicament et Sante*, STS Edition, Paris, 1991;47-58.

11 Mosimann JE, Wiseman CV, Edelman RE. Data fabrication: can people generate random digits? *Accountability in Research* 1995;4:31-55.

12 Preece DA. Distribution of final digits in data. *Statistician* 1981;30:31-60 .

13 Taylor R, McEntegart D, Stilman E. Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. *Drug Information Journal* 2002;36:115-125.

14 Evans S. Fraud and misconduct in medical science, in Armitage P. and Colton T. (eds), *Encyclopaedia of Biostatistics*, Wiley, Chichester, 1998;1583-1588.

15 Hill TP. A statistical derivation of the significant-digit law. *Statistics in Science* 1996;10:354-363.

16 Hill TP. The first-digit phenomenon. *Scientific American* 1998;86(4):358-363.

17 Danesh N, Kooshkghazi M. How does correlation structure differ between real and fabricated data-sets? *BMC Medical Research Methodology* 2003;3:18.

18 Chernoff H. the use of faces to present points in k-dimensional space graphically. *Journal of the American statistical Association* 1973;68:361-368.

19 Al-Marzouki S, Evans S, Marshall T, Roberts I. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *BMJ* 2005;331: 267-270.

CHAPTER 5

Statistical assessment of potentially fabricated data

5.1 Introduction

Most statistical analyses of clinical trials are undertaken on the presumption that the data are genuine. Large accidental errors can be detected during data analysis (1,2) but it might be assumed that falsification or fabrication of data would be done in a way that attempts to conceal its false nature, with any large discrepancies being avoided. The previous chapter showed that fraudulent data can have particular statistical features that are not evident in genuine data containing accidental errors and a number of analytic methods have been discussed to detect fraud in clinical trials (3,4).

Techniques among those discussed in Chapter 4 are applied in this chapter, as appropriate, to demonstrate fabrication in a set of data beyond reasonable doubt. In the literature, I have found no application of statistical techniques on a real dataset.

Data are examined from two randomised controlled trials using analytical and graphical techniques. In one trial, *BMJ* referees had raised the possibility of scientific misconduct based on inconsistencies in calculated P values compared with the means, standard deviations, and sample sizes presented (5). For comparison, a second trial for which there were no such concerns was analysed using the same methods.

5.2 Trial 1: The diet trial

The first trial, which will be referred to as "the diet trial", was a single blind randomised controlled trial of the effects of a diet and vegetable enriched diet in 831 patients with coronary heart disease, including patients with angina pectoris, myocardial infarction or surrogate risk factors. Study participants were stated to be randomly allocated to the intervention diet (Group I, N=415) or to the control group, which was the patient's usual diet (Group C, N=416). The aim was to examine the effect of the intervention diet on risk factors for coronary artery disease. After two years, according to dietary diaries, patients in Group I received a higher percentage of calories from complex carbohydrates, had a higher polyunsaturated to saturated fat ratio, a lower fat diet, a larger amount of soluble fibre and antioxidant nutrients, and a lower saturated fat and cholesterol than Group C. Because of the *BMJ* reviewers' suspicions about the integrity of the data, the *BMJ* requested the original trial data. These were provided by the first author on hand-written sheets, which were then computerised with appropriate checks to avoid transcription errors. The data are considered in the two randomised groups at baseline, Group I Group C. I do not present data from two years follow-up, because differences between groups could arise as a result of the intervention. The variables analyzed are shown in Table 5.1, some of which are the results of laboratory tests.

Table 5.1 Variables studied from the diet trial

Height	Protein
Weight	Fat
Diastolic blood pressure (DBP)	Saturated fat
Systolic blood pressure (SBP)	Fibre
Cholesterol	Soluble fibre
Fasting blood glucose	Caffeine
Total cholesterol	Salt
Triglycerides	Vitamin C
Energy	Carotene
Total carbohydrates	Vitamin E
Complex	Vitamin A

5.3 Trial 2: The drug trial

The second trial, which will be referred to as "the drug trial", was a randomised controlled trial of the effects of drug treatment in 21750 patients with mild hypertension from 31 centres. Five centres were randomly selected; these centres included 1047 patients, (centre 1, N= 271), (centre 2, N= 174), (centre 3, N= 193), (centre 4, N= 210), and (centre 5, N= 199), a number broadly comparable with the diet trial. Study participants were randomly allocated to receive the drug (Group I, N=509) or a placebo (Group C, N=538). The aim of the trial as a whole was to determine whether drug treatment reduced the occurrence of stroke, death due to hypertension and coronary events in men and women aged 35-64 years, when followed for 2 years. The drug trial data were provided by the trial investigators as computer files. The data were sorted by the date of randomization. The data are presented by centres at baseline and at two year follow up, (again data from the follow-up is not presented) using the same notation as for the diet trial. The variables used in this study were those available that were in common with the diet study. These are height, weight, systolic blood pressure (SBP), diastolic blood pressure (DBP), and cholesterol. Further details of the methods and results from that trial have been published (6).

5.4 Statistical methods

This section attempts to address the question "Are there characteristics of the data in the diet trial that are sufficiently "abnormal" to support the suspicion that they have been subject to misconduct, including the possibility of being fabricated or falsified"? It attempts to assess the strength of evidence in relation to this question.

It is believed that there was no misconduct in the drug trial, and that any unusual features of the data from this trial are no more than would be expected from data collected in good faith with reasonable attention to precision and accuracy. This should also show that some unusual patterns of data are compatible with them nevertheless being genuine. In that sense, the drug trial as a whole is used as a "Control".

5.4.1 Exploratory data analysis

The mean, median, mode, standard deviation (SD), the maximum and the minimum values in the two sub-groups of the diet and the drug data set are shown in Table 5.2.

Table 5.2 Mean, Median, Mode, SD, minimum and maximum for the two treatment groups at baseline in the two trials

	Diet		Drug	
	Intervention	Control	Intervention	Control
Height (cms)				
Mean	165	165	162	163
Median	165	165	161	163
Mode	165	165	160	157
SD	6.9	3.93	9.34	9.38
Min	140	140	138	140
Max	179	178	190	188
Weight (kg)				
Mean	65.74	65.59	69.95	70
Median	66	66	70	69
Mode	65	65	67	61
SD	7.89	7.64	11.54	12.35
Min	40	39	40	36
Max	87	85	111	120
Systolic blood pressure (mm Hg)				
Mean	134.2	131.9	184.3	184.8
Median	130	130	184	184
Mode	130	130	186	180
SD	18.47	16.91	12.45	13.05
Min	100	100	160	160
Max	200	195	209	210
Diastolic blood pressure (mmHg)				
Mean	86.5	86.7	91.6	91.5
Median	86	85	92	92
Mode	80	85	91	90
SD	9.98	9.17	10.99	11.56
Min	60	60	46	50
Max	112	120	114	117
Cholesterol (mmol/L)				
Mean	5.46	5.44	6.68	6.58
Median	5.48	5.48	6.6	6.5
Mode	5.43	5.43	5.9	6.1
SD	.351	.295	1.24	1.17
Min	4.53	2.95	3.6	3.7
Max	6.52	6	12	10.8

This table compares the two trials, and groups within the trials, for the five common variables. The first noticeable difference between the two trials is that generally standard deviations are smaller in the diet trial for the height and cholesterol measurements. This is only weak evidence of something unusual since trials can differ markedly in the participants included. What is also notable is that there tends to be a difference between groups at

baseline in the diet trial. This difference is sometimes in the average but also in the variability. While such between-group differences are reasonable post-treatment there should be no large differences between randomised groups within an RCT at baseline. Fabricating data, without the aid of a computer, to have a certain degree of variability is inherently more difficult.

5.4.2 Statistical tests

5.4.2.1 Comparison of means & variances between randomised groups at baseline

It is not recommended to carry out statistical significance tests at baseline in randomised trials, since in a truly randomised trial any baseline imbalance is simply due to chance. Some significant differences (about 1 in 20) will occur. However where misconduct is suspected, then such tests can be used to demonstrate misconduct. Even if there can be slight differences in means, it would be unusual to see more than 1 in 20 of these to be significant at $P<0.05$ and very rare to see several tests or single tests with extremely low P values. The same argument would apply to comparisons of the variability. Technically, the tests for equality of variability use the variance- the square of the standard deviation. Table 5.3 shows for each trial the results of statistical significance tests for differences in means and also in variances between the intervention and control groups at baseline for all variables. Statistically significant results are shaded in grey.

Table 5.3 Baseline comparison of the two intervention groups, diet trial and drug trial

	Diet trial				Drug trial			
	Levene's F-Test for Equality of Variances		t-test for Equality of Means		Levene's F-Test for Equality of Variances		t-test for Equality of Means	
	F	Significance	t	Significance (2-tailed)	F	Significance	t	Significance (2-tailed)
Height	71.15	1.4×10^{-16}	-.508	.612	.032	.858	.789	.431
Weight	.204	.652	.284	.776	2.493	.115	.157	.875
SBP	4.812	.029	1.885	.060	1.394	.238	.650	.516
DBP	4.366	.037	-.269	.788	.848	.357	-.136	.892
Cholesterol	28.77	1×10^{-7}	1.188	.235	1.664	.197	-1.293	.196
Fasting blood glucose	8.21	.004	-.574	.566				
Total cholesterol	.043	.835	-.347	.729				
Triglycerides	21.97	3×10^{-6}	.484	.628				
Energy	.983	.322	-1.565	.118				
Total carbohydrates	1.966	.161	.236	.814				
Complex	12.86	0.0004	14.757	6×10^{-44}				
Protein	15.18	0.0002	5.018	6×10^{-7}				
Fat	20.52	7×10^{-6}	-2.883	.004				
Saturated fat	15.21	0.0001	3.853	.0001				
Fibre	94.23	4×10^{-21}	-8.471	2×10^{-16}				
Soluble fibre	10.13	.002	-6.950	7×10^{-12}				
Caffeine	2.407	.121	.957	.339				
Salt	39.72	5×10^{-10}	-.377	.706				
Vitamin C	.007	.931	-5.617	3×10^{-8}				
Carotene	51.06	2×10^{-12}	29.830	2×10^{-133}				
Vitamin E	25.71	5×10^{-7}	5.907	5×10^{-9}				
Vitamin A	51.42	2×10^{-12}	4.488	8×10^{-6}				

In the drug trial, none of the baseline means and none of the baseline variances showed statistically significant differences between the two groups. In contrast, the diet trial shows highly significant differences in variances for 16 of the 22 variables and highly significant differences in means for 10 variables. Several of these P values are quite extraordinarily small. This is simply very implausible in the context of genuine data from a randomised trial. The expectation is that about 5% of such comparisons would have $P<0.05$, and extremely small P values should not occur. The differences in means might occur if there had been some subversion of the randomization process to obtain a desired result. It is very unlikely indeed that this would translate into a difference in variance in the way that has happened here. The extent of this non-homogeneity between the two randomised groups for so many different factors seems to be explicable only if the data were invented using different people to invent the intervention and control groups.

5.4.2.2 Chi squared test of the final digit

This chi-squared test can be used to test goodness-of-fit to a hypothesised distribution. The final digit of any measurement such as blood pressure will reflect the measurement process. Checking digit preference, especially terminal digit preference that may be expected to have a uniform distribution (see section 4.3.1).

The null hypothesis is that the data come from a uniform distribution – the last digits are all equally likely. The calculated values of χ^2, with their associated probability (P) are shown in Table 5.4 for all variables for the two datasets.

Table 5.4 χ^2 value (with P value) for the final digit at the baseline in the diet and drug trials

	Diet trial		Drug trial	
	Chi-square (P)		Chi-square (P)	
	Intervention	Control	Intervention	Control
Height			289.6 (4×10^{-57})	297.5(9×10^{-59})
Weight	128 (4×10^{-23})	23 (0.00655)	4.654 (.863)	6.796 (.658)
SBP	1796 (U)	1470 (U)	6.972 (.640)	12.929 (.166)
DBP	763 (2×10^{-158})	820 (9.7×10^{-171})	12.669 (.178)	15.160 (.087)
Cholesterol	554 (2×10^{-113})	430 (6×10^{-87})	14.756 (.098)	5.306 (.807)
Fasting blood glucose	478 (4×10^{-97})	538 (5×10^{-110})		
Total cholesterol	1053 (6×10^{-221})	1522 (U)		
Triglycerides	642 (2×10^{-132})	963 (2×10^{-201})		
Energy	2151 (U)	2630 (U)		
Total carbohydrates	207 (1×10^{-39})	927 (7×10^{-194})		
Complex	231 (1×10^{-44})	939 (3×10^{-196})		
Protein	54 (2×10^{-8})	251 (5×10^{-49})		
Fat	229 (2×10^{-44})	437 (2×10^{-88})		
Saturated fat	123 (4×10^{-22})	98 (4×10^{-17})		
Fibre	263 (2×10^{-51})	1127 (9×10^{-237})		
Soluble fibre	273 (1×10^{-53})	1086 (6×10^{-228})		
Caffeine	613 (3×10^{-126})	694 (1×10^{-143})		
Salt	288 (9×10^{-57})	301 (2×10^{-59})		
Vitamin C	304 (5×10^{-60})	411 (6×10^{-83})		
Carotene	1470 (U)	1156 (5×10^{-243})		
Vitamin E	118 (3×10^{-21})	101 (8×10^{-18})		
Vitamin A	705 (6×10^{-146})	799 (3×10^{-166})		

(Chi-squared has 9 degrees of freedom)
U means that the p value is too small for calculation.

The chi-squared values are highly statistically significant for height (which therefore shows strong digit preference) but not for any of the other measures in the drug trial, so there is no evidence of digit preference for these other variables. For research purposes, to avoid digit preference, a blood pressure measuring device known as a "Hawksley Random Zero Sphygmomanometer" is often used, and was used in the drug trial. Thus, it is not surprising

that this trial has no digit preference for blood pressure. In the diet trial, the digit preference for blood pressure is marked, but this is definitely not evidence that any misconduct has occurred. All of the chi-squared values are highly significant in this trial. This is not surprising for blood pressure or possibly also weight, but is not expected for a lab test such as cholesterol.

5.4.2.3 Chi squared test to compare the distribution of the final digit between randomised groups

This test is for any difference in the distributions of final digit between the two randomised groups in each trial. This is a test as to whether digit preference is the same in the two groups created by randomization. If randomization has been properly carried out on real data, then the null hypothesis for this test must by definition be true. This means that, even if there is digit preference, this preference should be very similar in the two groups. In the drug trial, Table 5.5, there are no significant differences between the two randomised groups in terms of the final digit, even in height where there was marked digit preference in the measure itself. However, for the diet trial, the final digit distributions are significantly different between the intervention and the control group at baseline for all variables, except for cholesterol, fasting blood glucose, energy, saturated fat, caffeine, carotene, and vitamin A.

Table 5.5 χ^2 value (with P value) for the final digit at the baseline in the diet and drug trials between the two randomised groups

	Diet trial		Drug trial	
	Chi-square (P)	df.	Chi-square (P)	df.
Height			3.44 (.944)	9
Weight	36 (3×10^{-5})	9	9.24 (.418)	9
SBP	26 (0.00019)	6	9.15 (.423)	9
DBP	16 (0.046)	8	11.5 (.243)	9
Cholesterol	13 (0.182)	9	4.95 (.838)	9
Fasting blood glucose	12 (0.2)	9		
Total cholesterol	46 (5×10^{-7})	9		
Triglycerides	48 (3×10^{-7})	9		
Energy	16 (0.064)	9		
Total carbohydrates	154 (2×10^{-28})	9		
Complex	135 (1.4×10^{-24})	9		
Protein	43 (2×10^{-6})	9		
Fat	40 (6.4×10^{-6})	9		
Saturated fat	15 (0.08)	9		
Fibre	157 (8×10^{-30})	8		
Soluble fibre	175 (6.5×10^{-33})	9		
Caffeine	15 (0.059)	8		
Salt	28.5 (0.001)	9		
Vitamin C	18 (0.03)	9		
Carotene	10 (0.266)	8		
Vitamin E	20 (0.017)	9		
Vitamin A	9.5 (0.4)	9		

The degrees of freedom are less than 9 when one or more digits do not appear

These patterns in the data are indicative of a very serious problem. The drug trial results show that even where digit preference occurs this is very similar between the groups. In the diet trial, both the pattern of digit preference (digit preference occurring in lab test results) and the fact that the digit preference pattern is markedly different between the groups formed by randomization make it clear that the data are not set out in a way that is

compatible with a truly randomised trial using real data. It is strong evidence of some form of misconduct.

5.4.2.4 Test of runs above and below the median

The runs test is a non-parametric, distribution-free test, which, as applied here, tests for consecutive values being randomly above or below the median. It is used to see if successive values are related to one another. This often happens in economic data, but is rarer in medical data. However, medical data can be subject to seasonal or other time-based fluctuations so this test is not as clear evidence of misconduct as some other tests. What is expected is that successive patients recorded as arriving in a trial will have random fluctuations in the values of most variables. In carrying out the test, it is also assumed that subjects are entered into the trial database either in their order of arrival and in some other order not related to the value of any of the variables under consideration. This appears to be a quite reasonable assumption as part of the null hypotheses. The values for the test and the associated P value are shown in Table 5.6 for the diet and drug trials.

Table 5.6 Runs test value (with p value) for all measures, at the baseline in the diet and drug trials

	Diet trial		**Drug trial**				
	Intervention	Control	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
Height			-1.67 (.096)	-4.3 (2×10^{-5})	-.4 (.689)	-1.38 (.168)	.43 (.67)
Weight	-5.9 (2×10-9)	-7.16 (8×10^{-13}	-1.4 (.162)	-.29 (.766)	.697 (.49)	-2.49 (.013)	-.204 (.84)
SBP	-9.1 (1×1019)	-7.96 (2×10^{-15})	-1.5 (.131	.64 (.524)	-1.1(.29)	.575 (.565)	-.06 (.95)
DBP	-6.9 (5×10^{-12})	-4.27 (2×10^{-5})	-2.4 (.017)	.78 (.437)	-.58 (.56)	-1.08 (.282)	.93 (.35)
Cholesterol	-9.98 (2×10^{-23})	-8.64 (6×10^{-18})	-.78 (.435)	.73 (.469)	-3(.005)	1.04 (.299)	-1.2 (.23)
Fasting blood glucose	-6.5 (7×10^{-11})	-8.44 (3×10^{-17})					
Total cholesterol	-5.4 (8×10^{-8})	-7.9 (3×10^{-15})					
Triglycerides	-6.8 (9×10^{-12})	-6.5 (6×10^{-11})					
Energy	-5.9 (5×10^{-9})	-7.1 (1×10^{-12})					
Total carbohydrates	-7.1 (1×10^{-12})	-8.3 (8×10^{-17})					
Complex	-9.1 (1×10^{-19})	-7.4 (2×10^{-13})					
Protein	-5.2 (2×10^{-7})	-5.3 (1×10^{-7})					
Fat	-7.6 (3×10^{-14})	-8.6 (6×10^{-18})					
Saturated fat	-7.7 (1×10^{-14})	-8.9 (4×10^{-19})					
Fibre	-9.98 (2×10^{-23})	-6.5 (8×10^{-11})					
Soluble fibre	-8 (9.9×10^{-17})	-9.3 (2×10^{-20})					
Caffeine	-4.8 (1×10^{-6})	-10.3 (8×10^{-25})					
Salt	-8 ((9.9×10^{-17})	-3.7 (0.00025)					
Vitamin C	-9.24 (2×10^{-20})	-9.4 (8×10^{-21})					
Carotene	-11.8 (2×10^{-32})	-8.3 (9×10^{-17})					
Vitamin E	-9.48 (3×10^{-21})	-6.8 (1×10^{-11})					
Vitamin A	-9 (2×10^{-19})	-6.9 (3×10^{-12})					

Table 5.6 shows the results from the runs test indicating that the diet trial contains, in virtually all instances, strong non-random sequences in the data. This test uses the trial data in the numbered order in which it was provided to the BMJ, and counts how many successive values are all either above or below the median. The data have not been sorted by any of the variables, and whilst it would be possible for perhaps cholesterol to reflect at baseline some serial relationship perhaps because diet were to be changing over time, this would not be expected for baseline values of weight or height. There was some evidence of non-random sequences for diastolic blood pressure in centre 1, height in centre 2, cholesterol in centre 3, and weight in centre 4 in the drug trial.

5.4.3 Graphical techniques of data exploration

The manual process of generating fraudulent data is not likely to produce randomness from one observation to the next. Four plots are now presented that are sensitive to such non-randomness. It should be noted that these do not, by themselves, prove randomness or non-randomness, but they can highlight patterns, outliers and relationships that appear to be non-random. It is argued that any tendency for connection between observations for consecutive participants in the data is unlikely to occur in genuine data, since there is no inherent connection of one case to the next. In fabricating data, it is likely that the fabricator will be unconsciously influenced in choosing each value by the values that have gone before. In addition, the bivariate plot is also presented to provide a graphical display of the relationship between two variables.

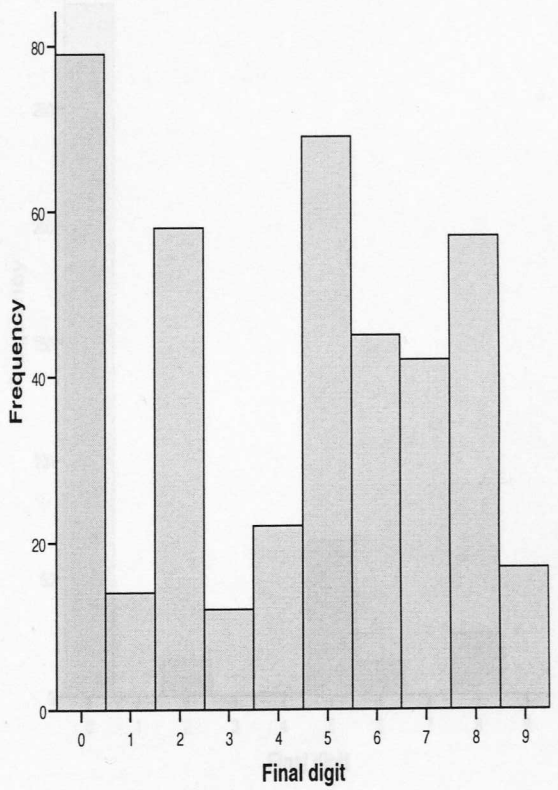
5.4.3.1 The histogram for final digit

The histogram plots for the final digits that occur in the measurements of each variable are shown in Figure 5.1a. One expects there to be an even distribution of the 10 digits, although preference for zero and five is common and does not necessary indicate any malpractice.

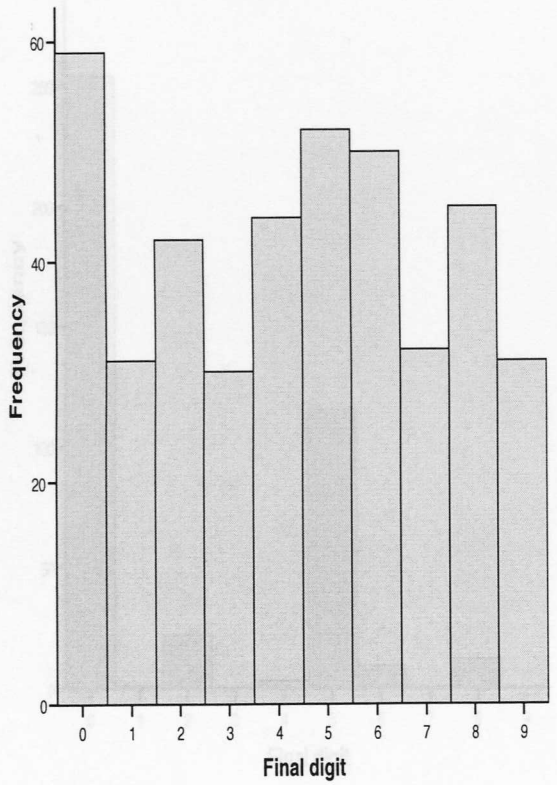
Figure 5.1a Histogram plots for both intervention and control groups at baseline in the diet trial

Weight

Intervention group

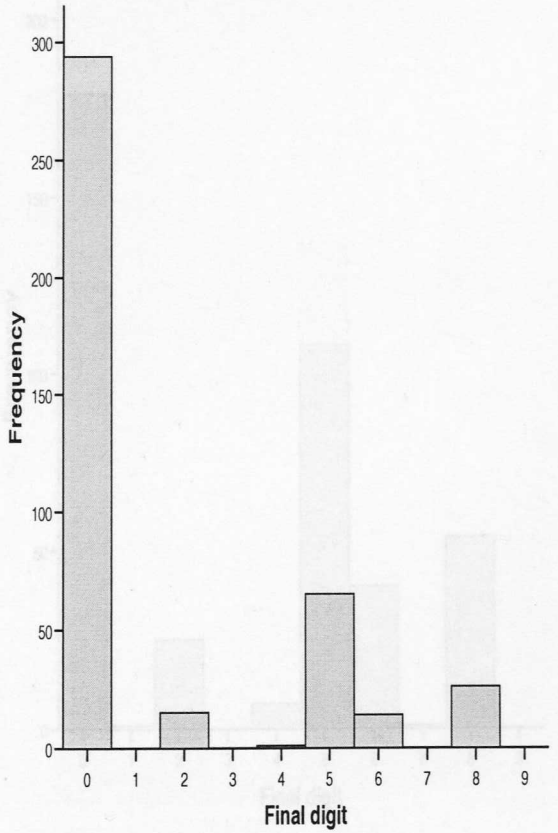


Control group

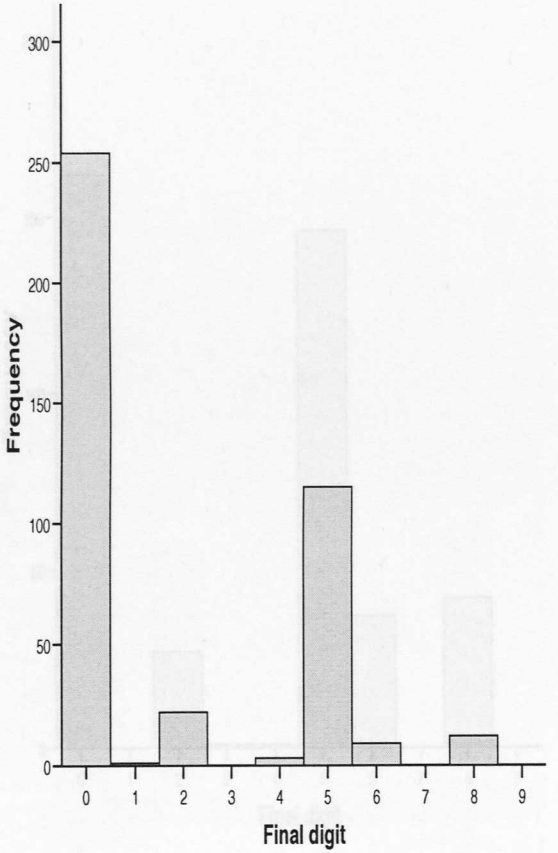


Systolic blood pressure

Intervention group

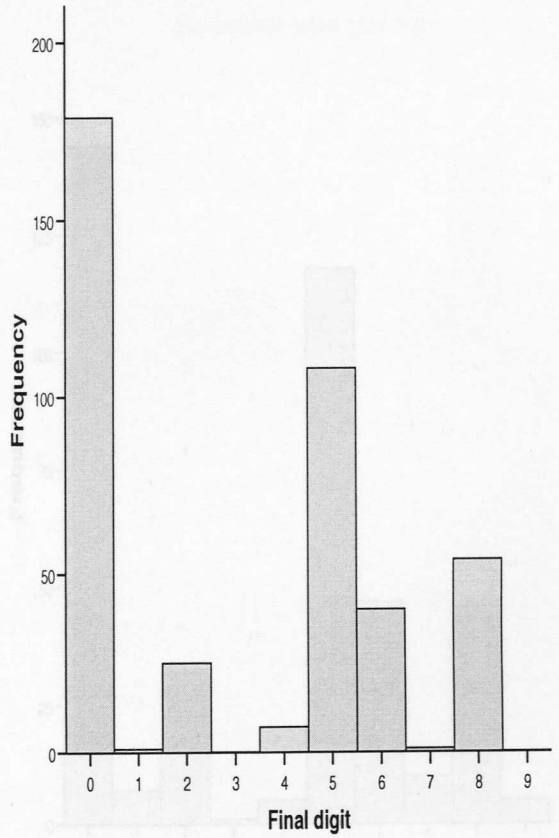


Control group

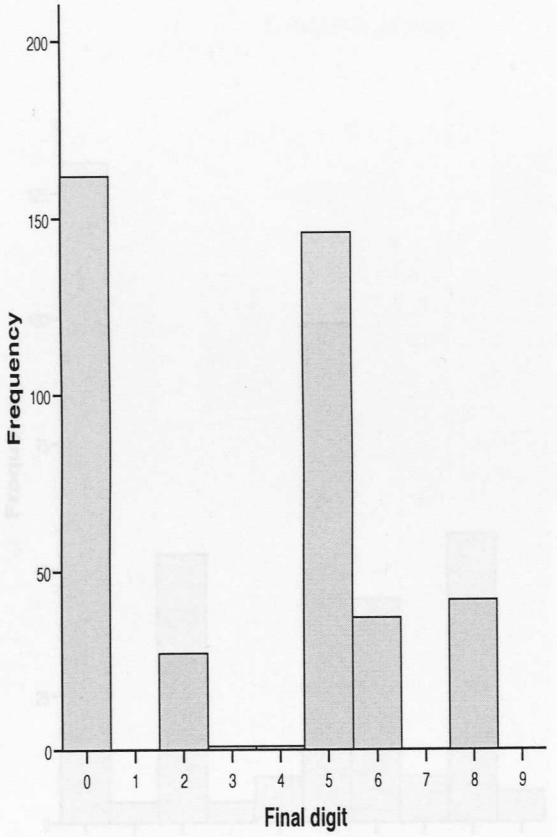


Diastolic blood pressure

Intervention group

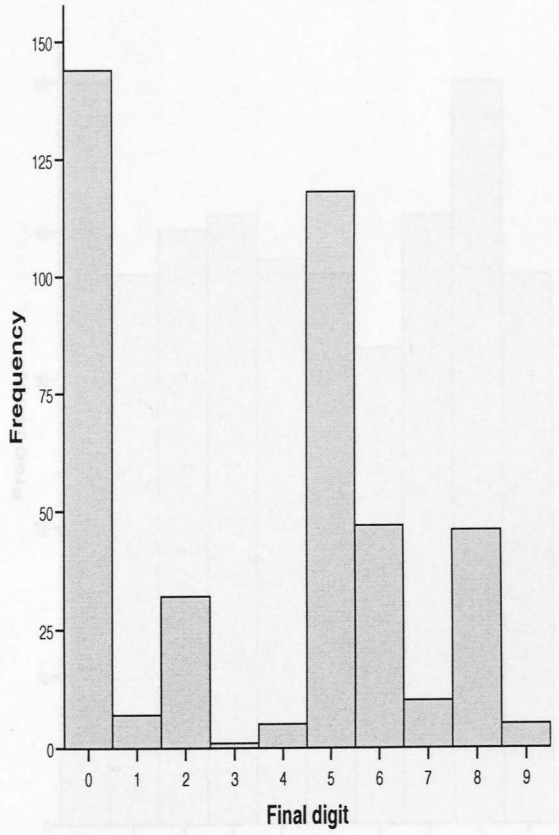


Control group



Cholesterol

Intervention group



Control group

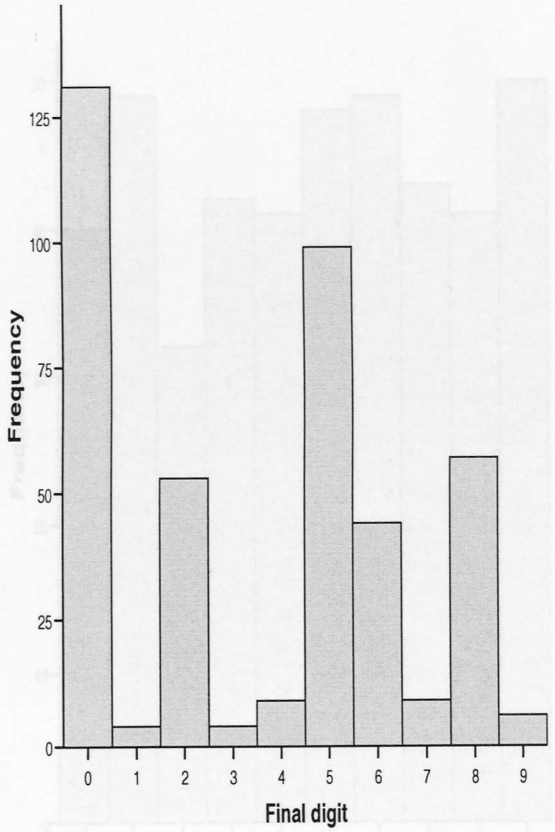
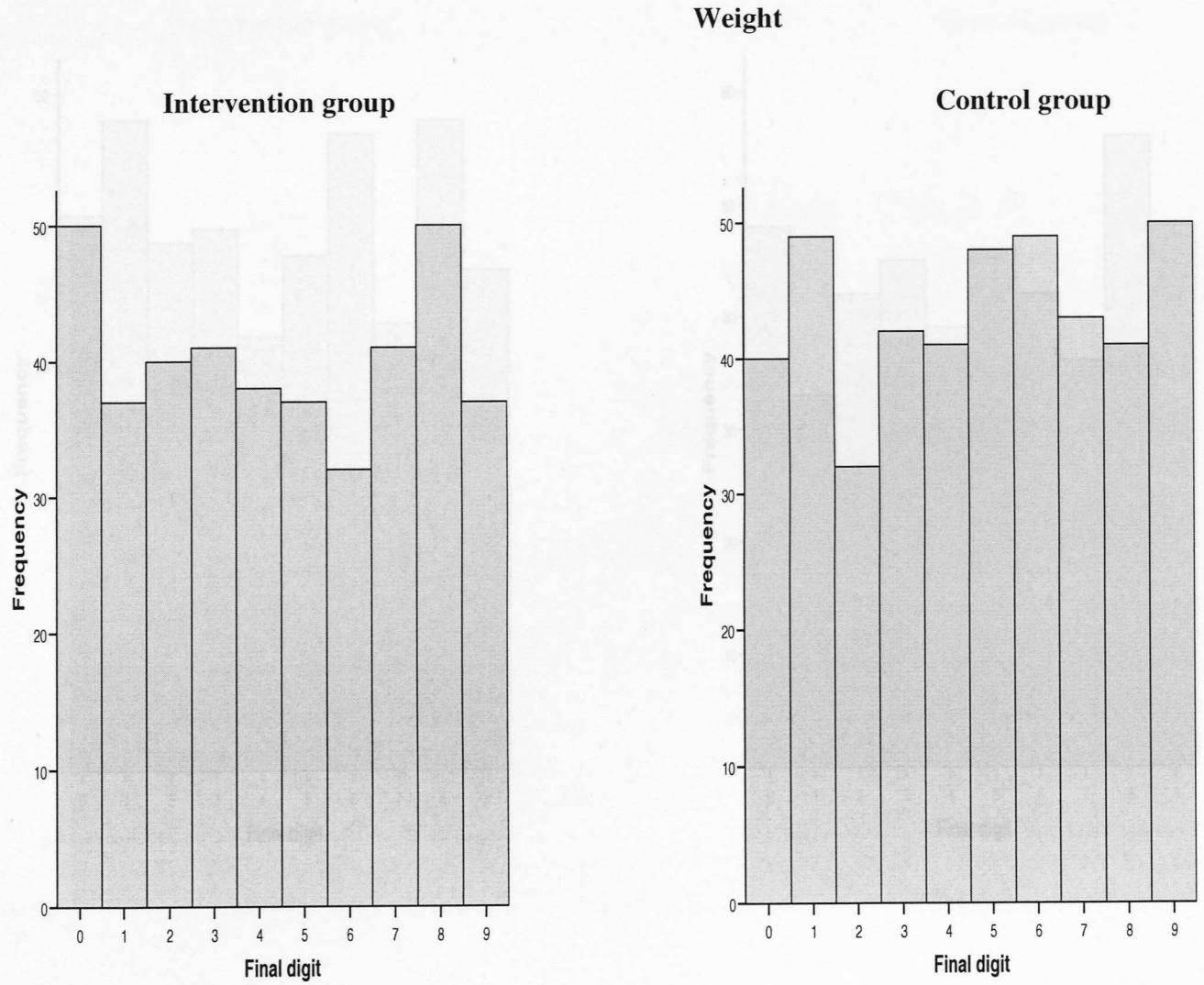
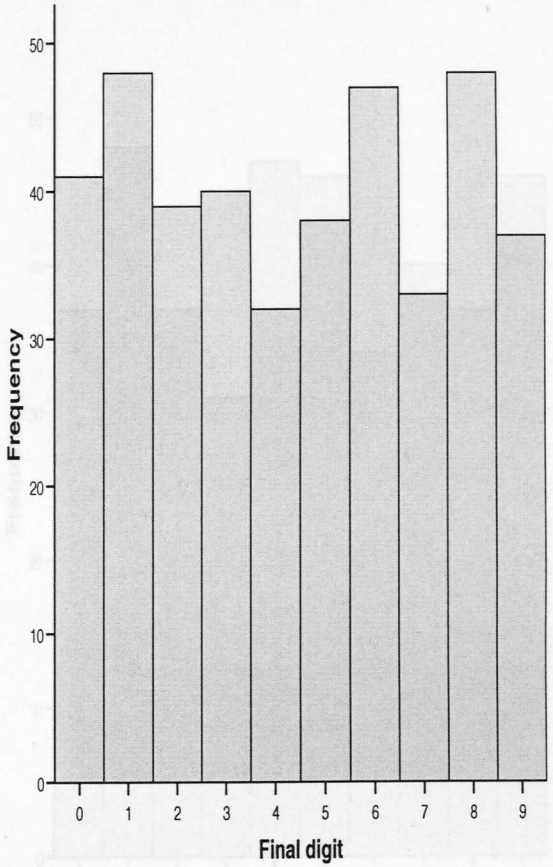


Figure 5.1b Histogram plots for both intervention and control groups at baseline in the drug trial

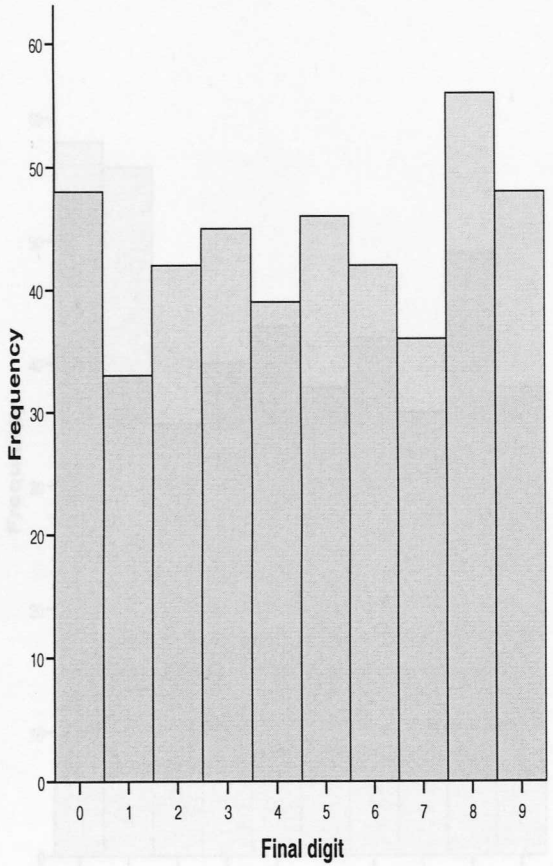


Systolic blood pressure

Intervention group

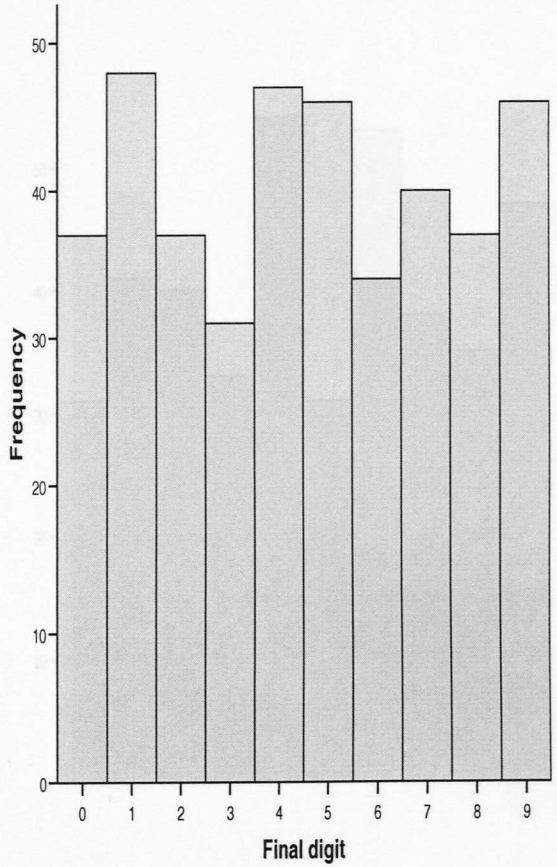


Control group

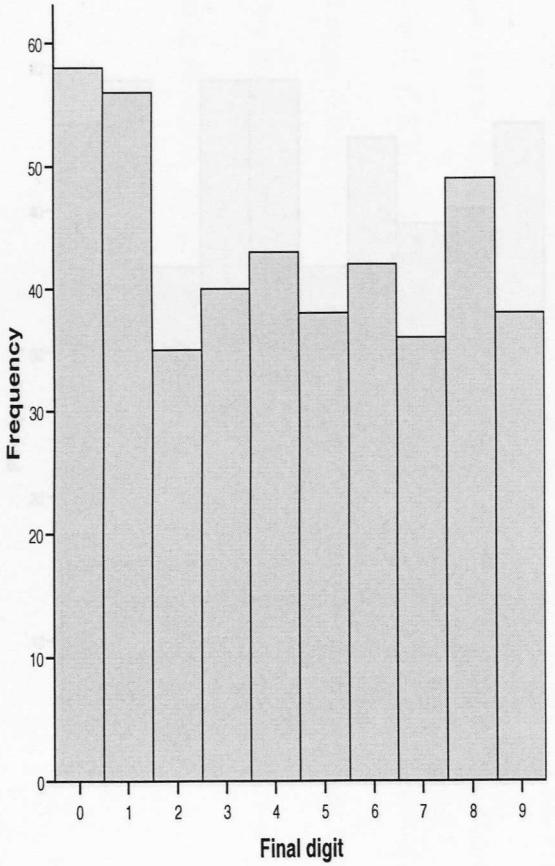


Diastolic blood pressure

Intervention group

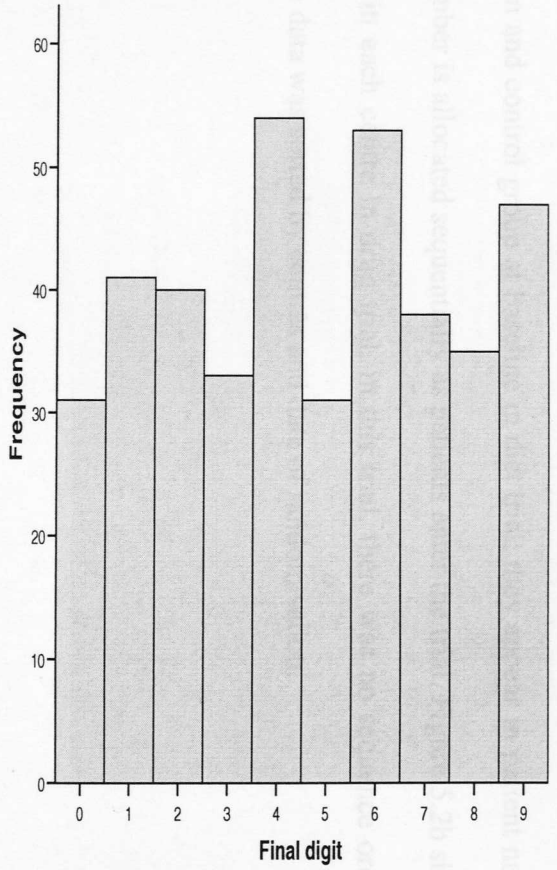


Control group

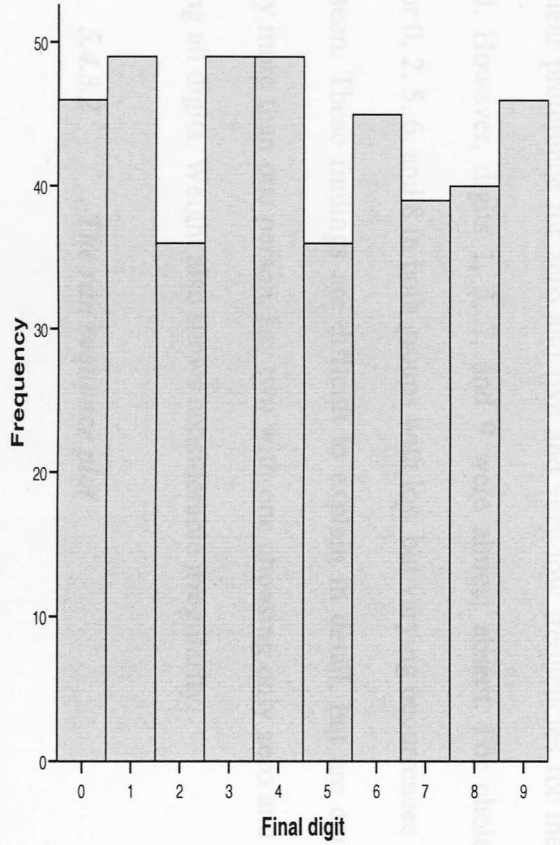


Cholesterol

Intervention group



Control group

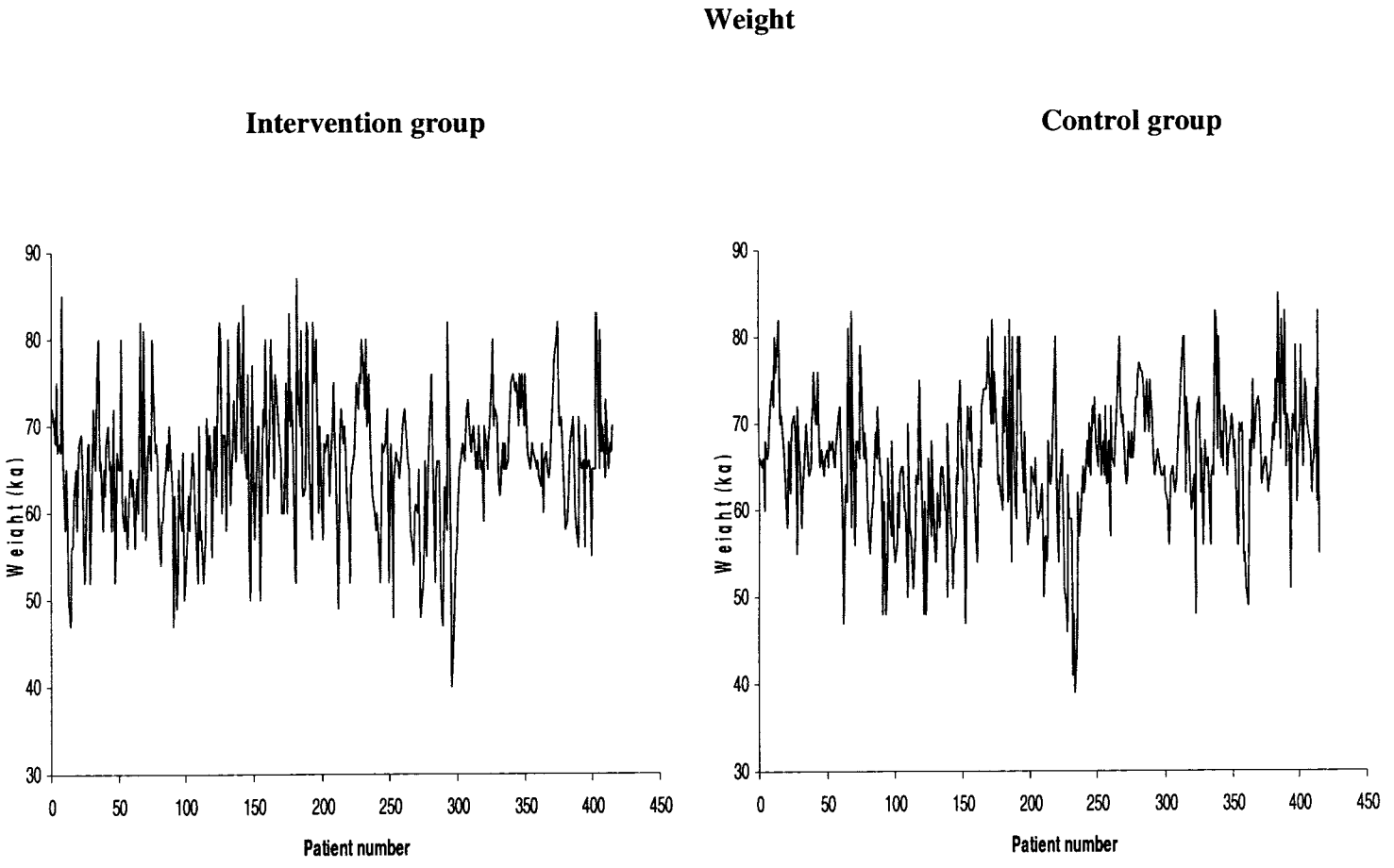


The histograms for the drug trial do not show any particular pattern. Chi-squared tests of equal numbers of each digit were presented in section 5.4.2.2 and were non-significant. For the diet trial, the results are very different. For weight in the intervention group, digit 1, 3, 4 and 9 are appeared with low frequencies. In systolic blood pressure and diastolic blood pressure, a clear preference for zero and five and very low frequencies for the even digits were noticed. However, digits 1, 3, 7, and 9 were almost absent. For cholesterol, digit preference for 0, 2, 5, 6, and 8 in both groups with low but varying occurrences for the other digits were seen. These findings are difficult to explain in detail, but are consistent with fabrication by more than one person, say two with one choosing only zero and five and the other choosing all digits. Weight also shows considerable irregularity.

5.4.3.2 The run sequence plot

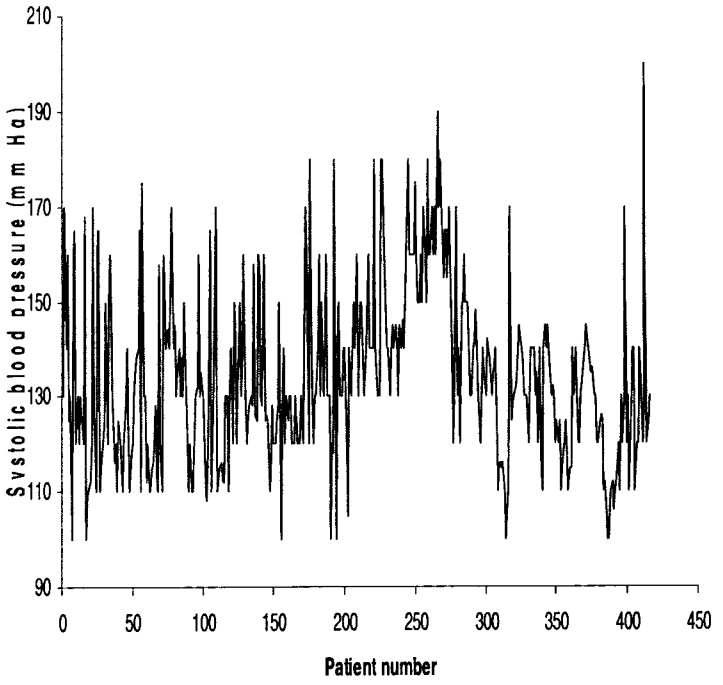
The run sequence plot is a graph of each observation against its position in the sequence of the data values. It is used to show a random pattern in the data; it is similar to and could give the same conclusions as the run test. Figure 5.2a, below shows the run sequences for the intervention and control group at baseline in diet trial; they appear in patient number order. Patient number is allocated sequentially as patients enter the trial. Figure 5.2b shows the run sequences in each centre in drug trial, in this trial, there was no sequence order for each patient; the data was sorted by centres and date of randomisation.

Figure 5.2a Run sequence plots for both intervention and control groups at baseline in the diet trial

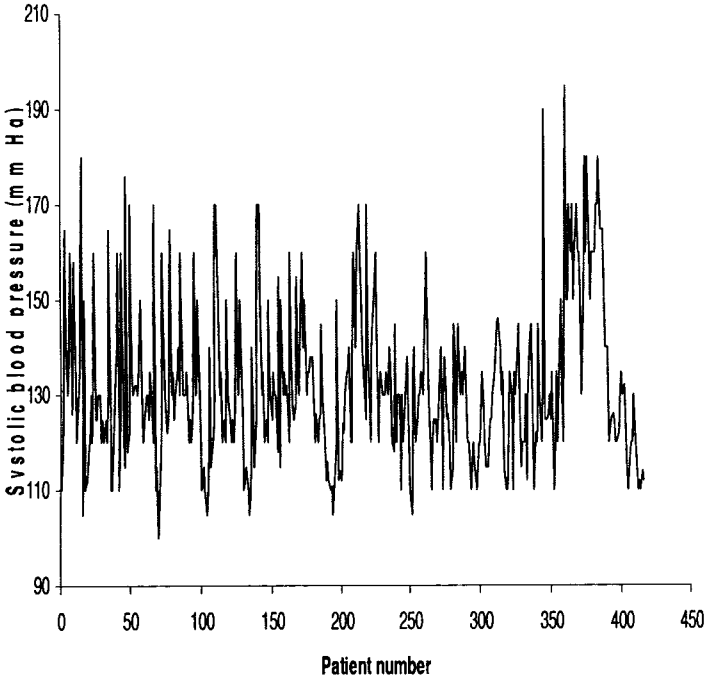


Systolic blood pressure (mm Hg)

Intervention group

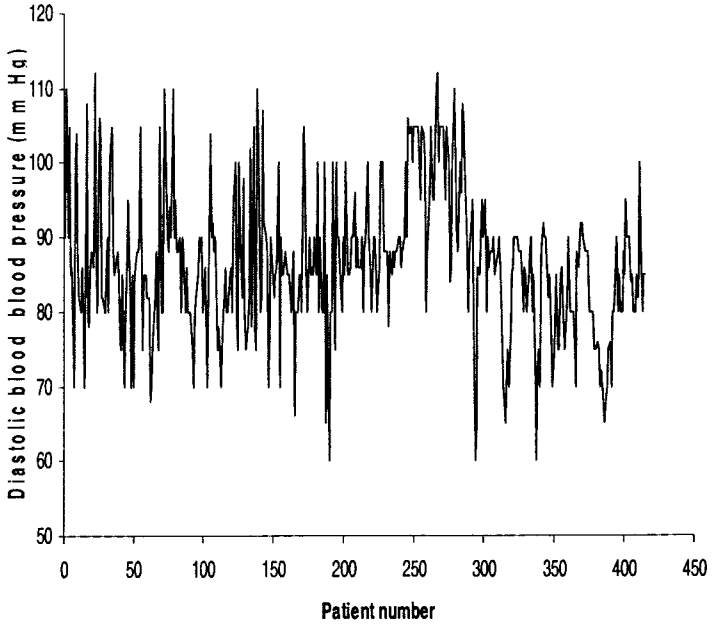


Control group

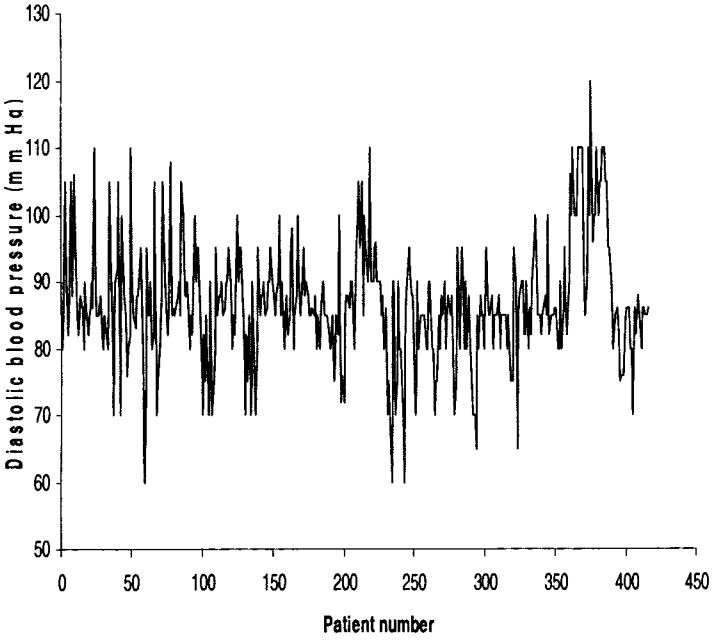


Diastolic blood pressure (mmHg)

Intervention group

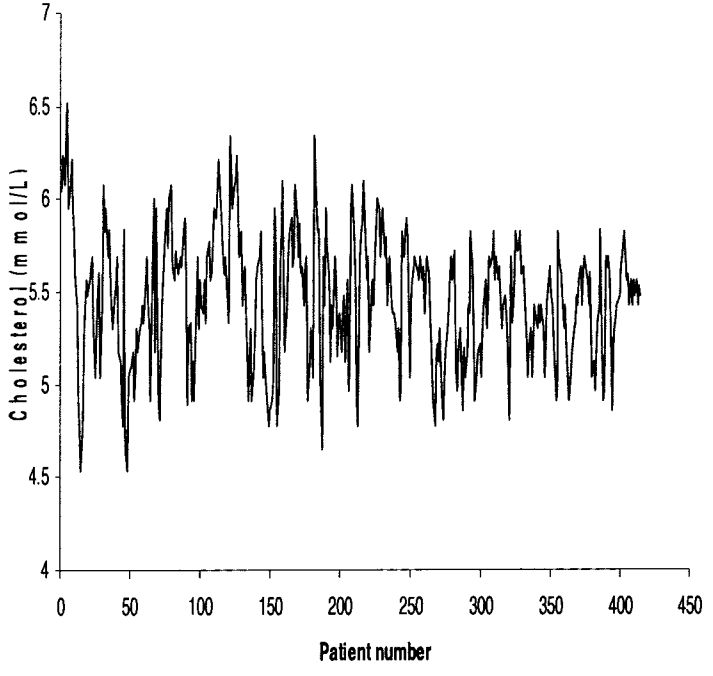


Control group



Cholesterol (mmol/L)

Intervention group



Control group

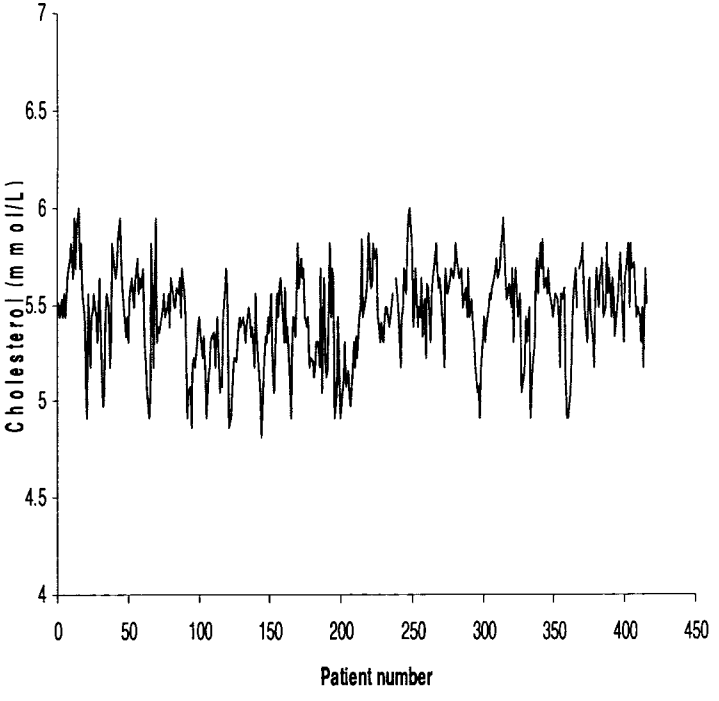
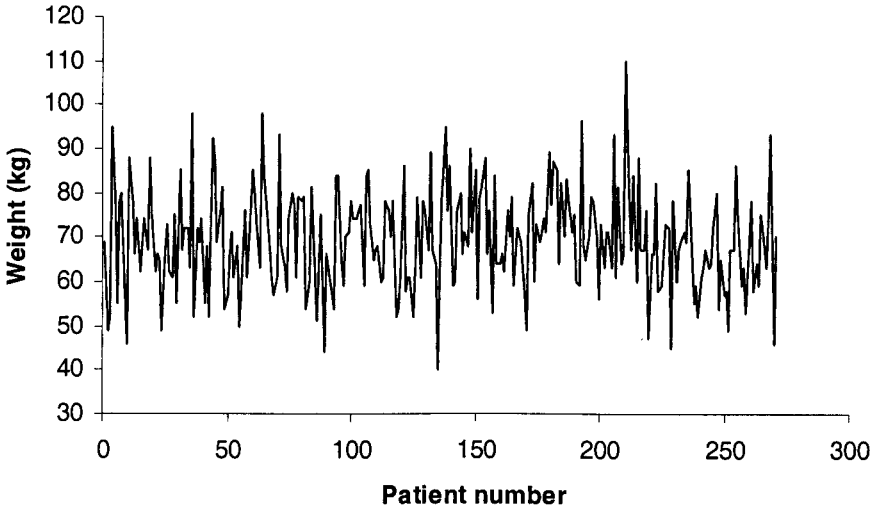
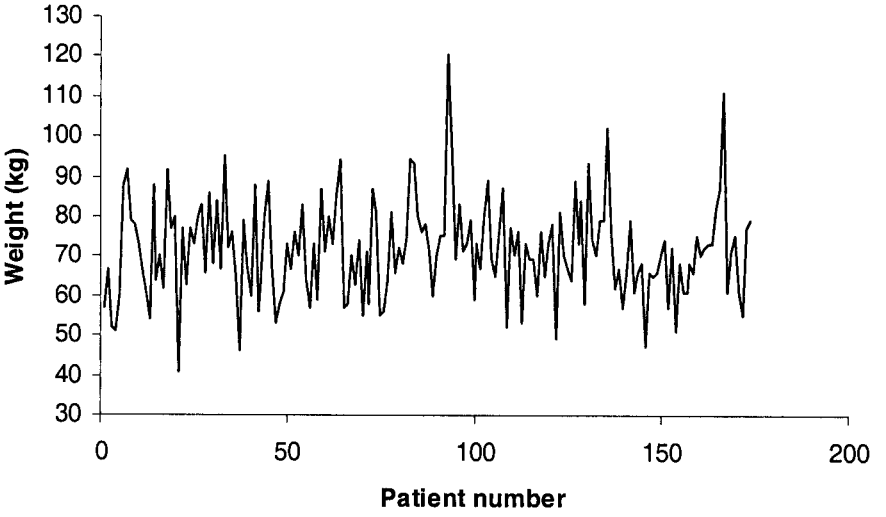


Figure 5.2b Run sequence plots for each centre at baseline in the drug trial

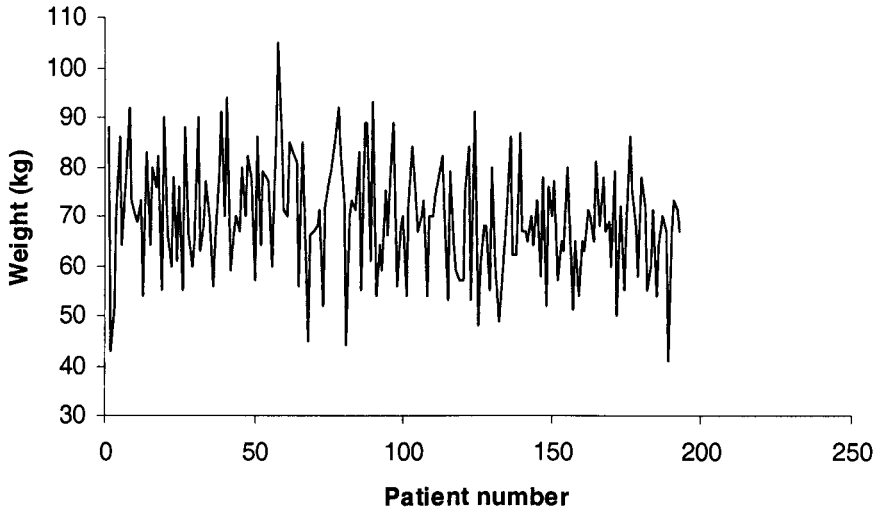
Centre 1



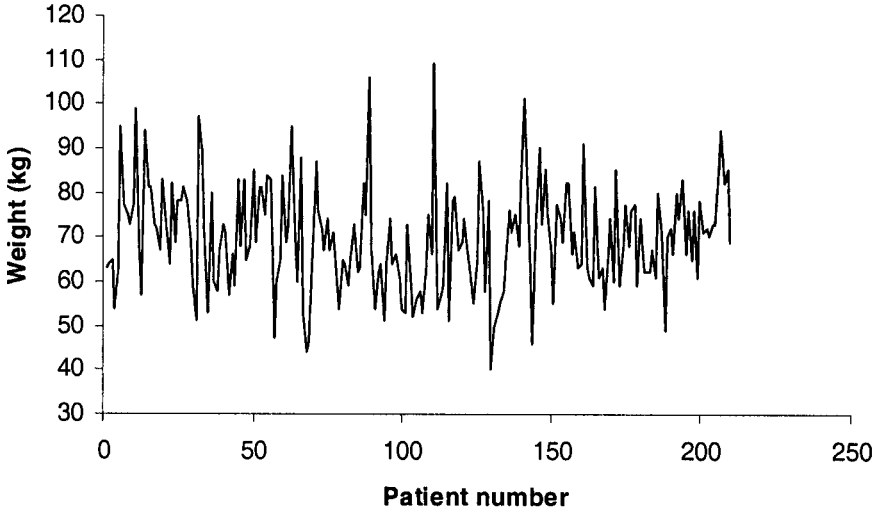
Centre 2



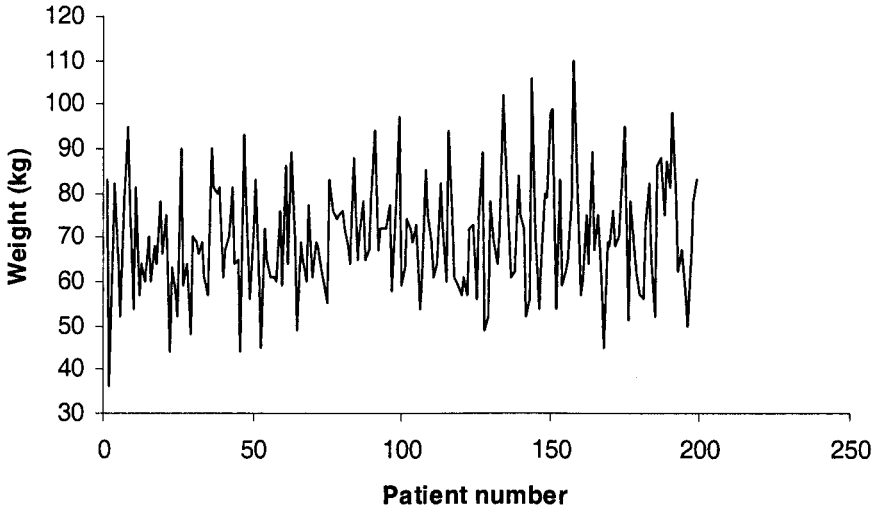
Centre 3



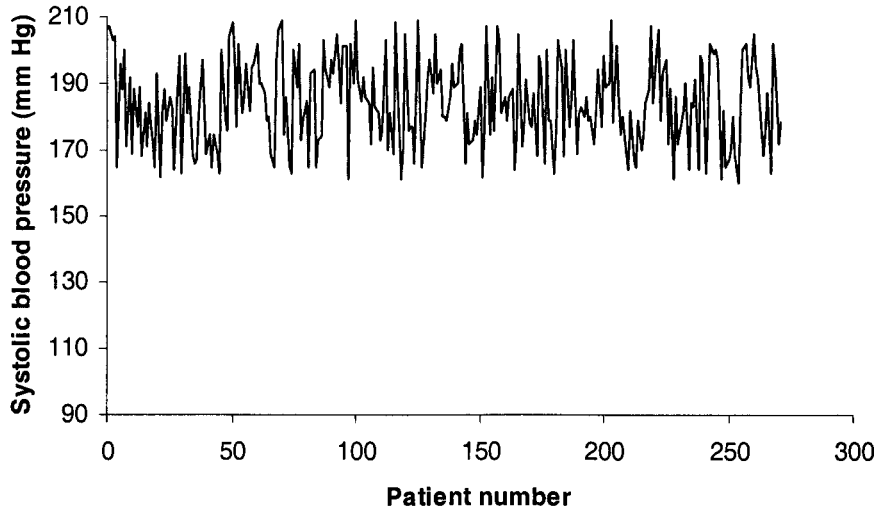
Centre 4



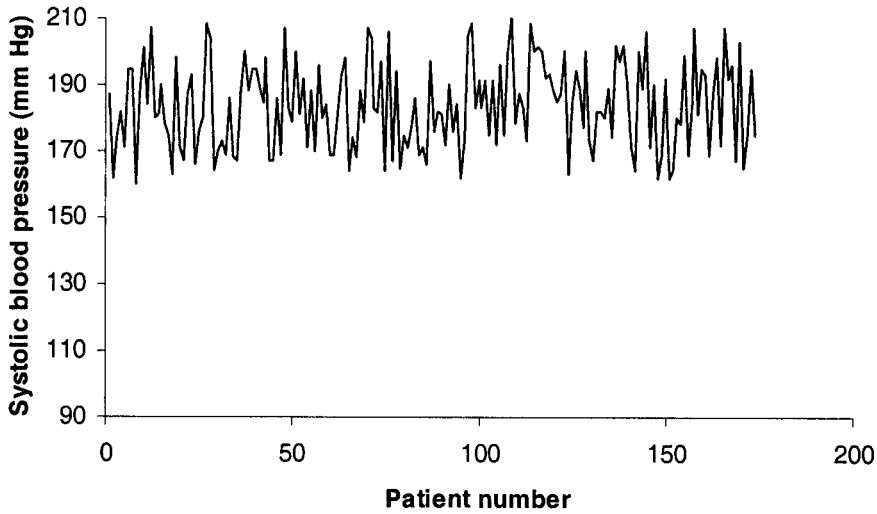
Centre 5


120
110
100
90
80
70
60
50
40
30
Weight (kg)
0
50
100
150
200
250
Patient number

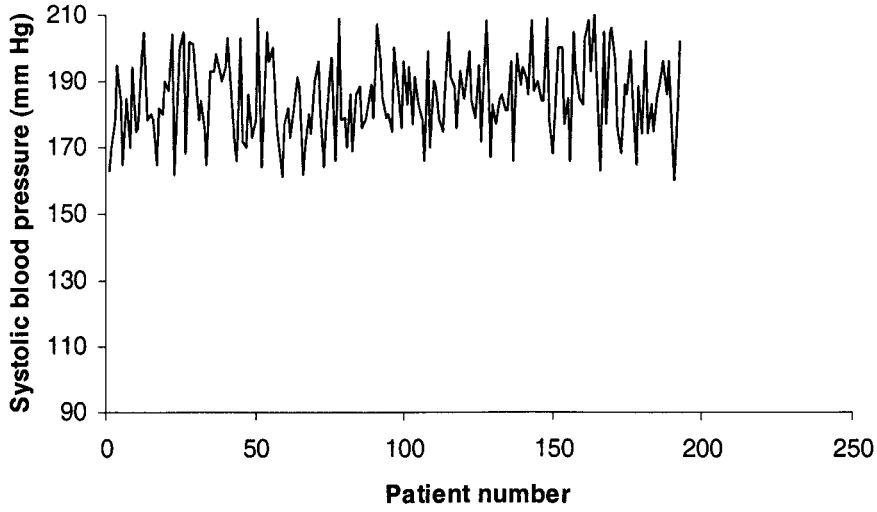
Centre 1



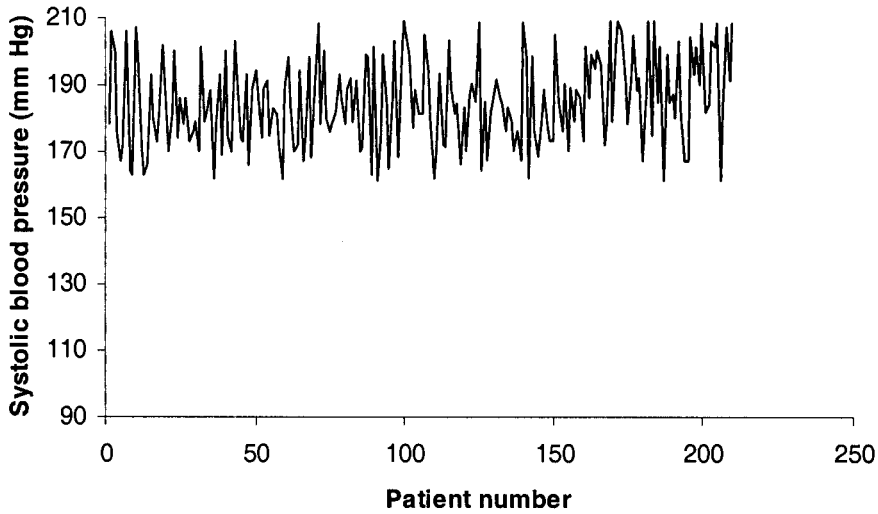
Centre 2



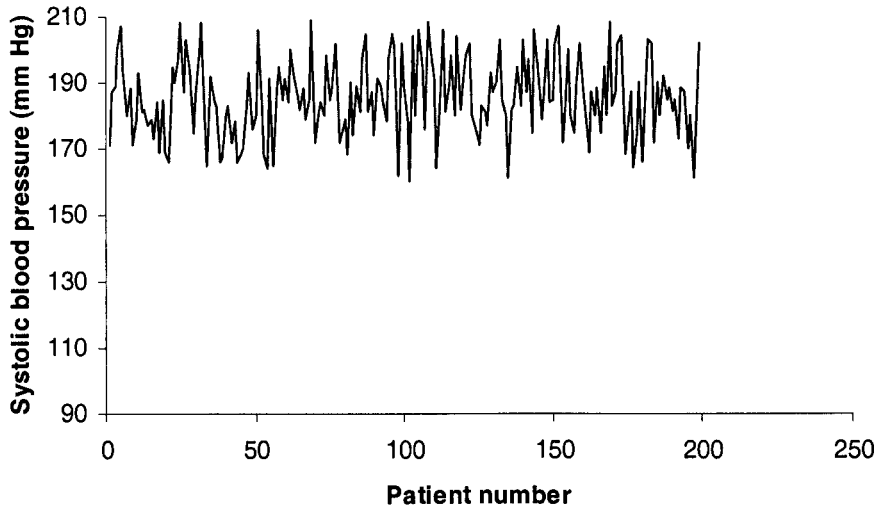
Centre 3



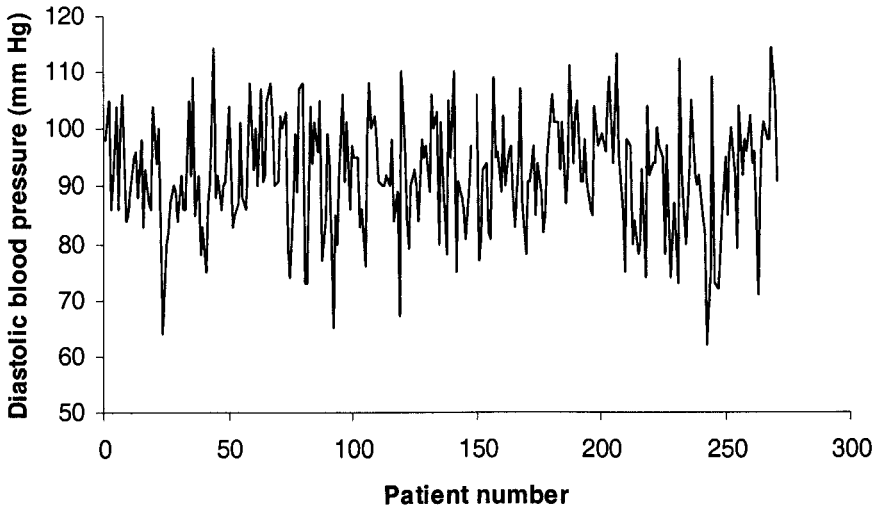
Centre 4



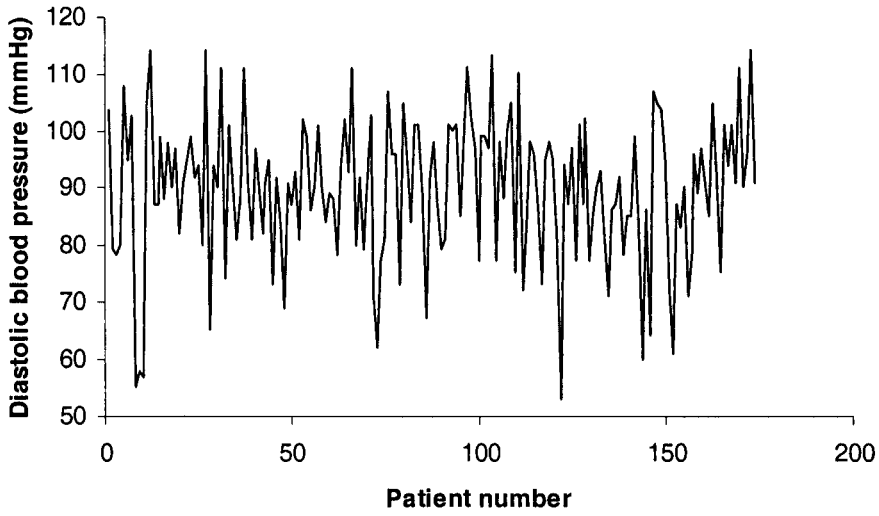
Centre 5


Systolic blood pressure (mm Hg)
210
190
170
150
130
110
90
0
50
100
150
200
250
Patient number

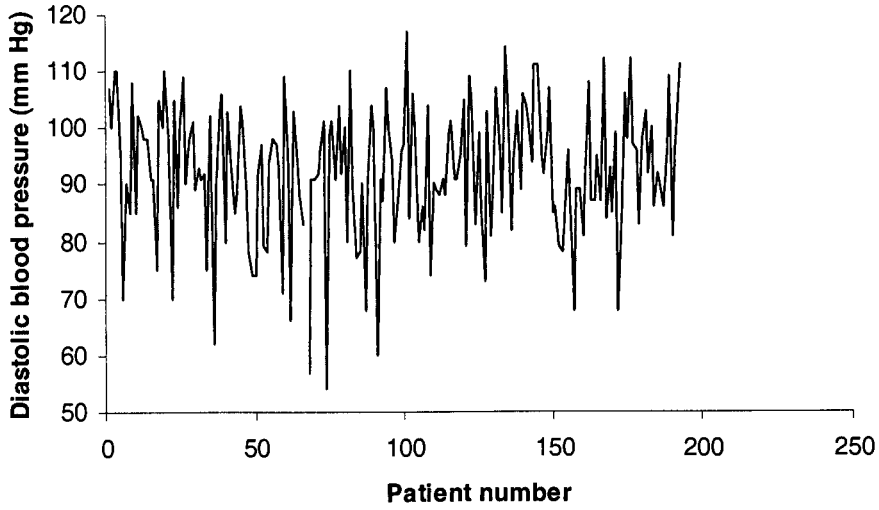
Centre 1



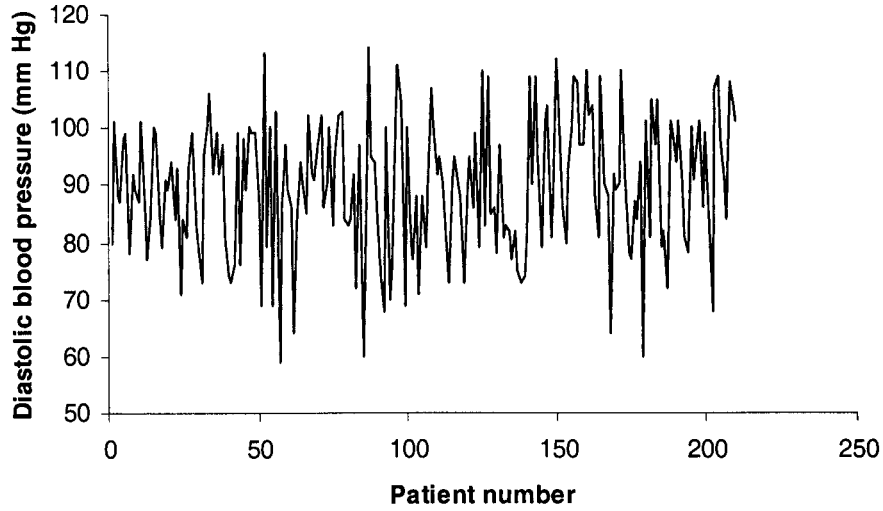
Centre 2



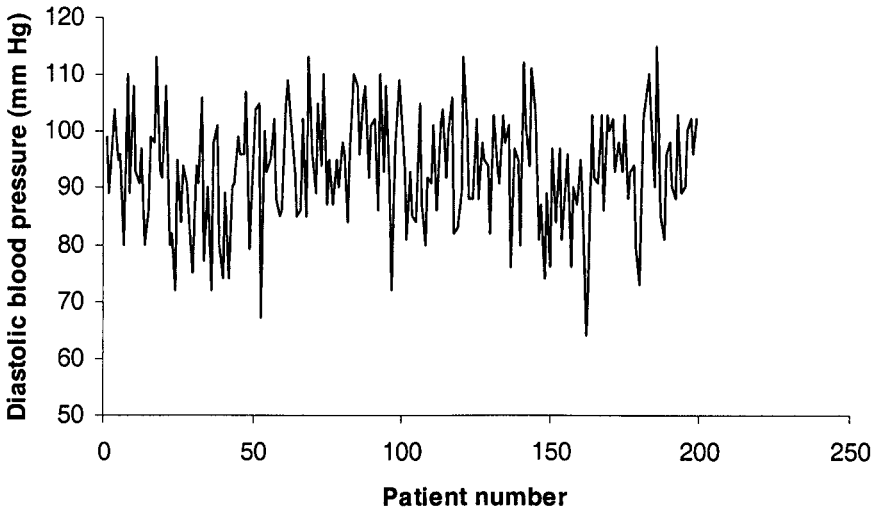
Centre 3



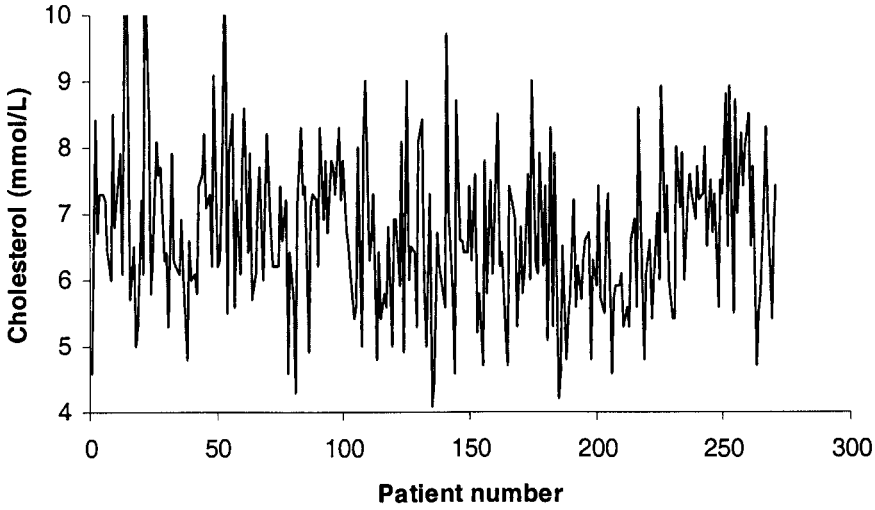
Centre 4



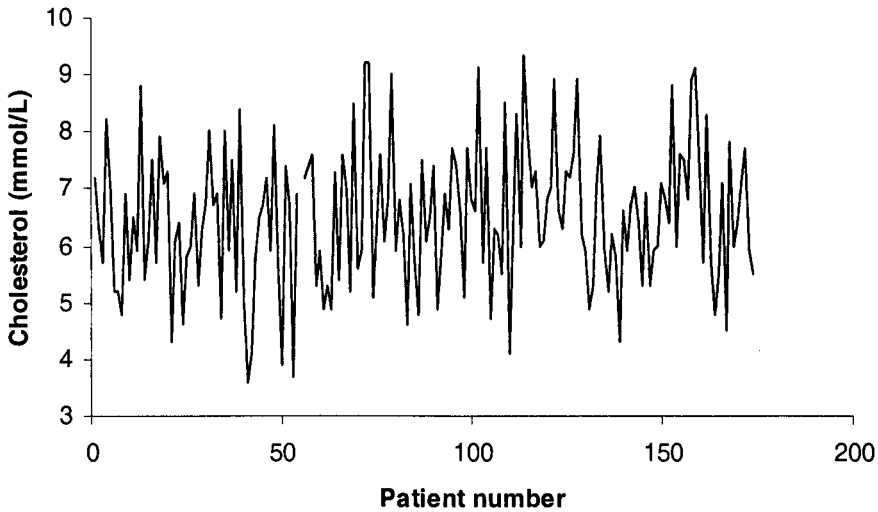
Centre 5


Diastolic blood pressure (mm Hg)
120
110
100
90
80
70
60
50
0
50
100
150
200
250
Patient number

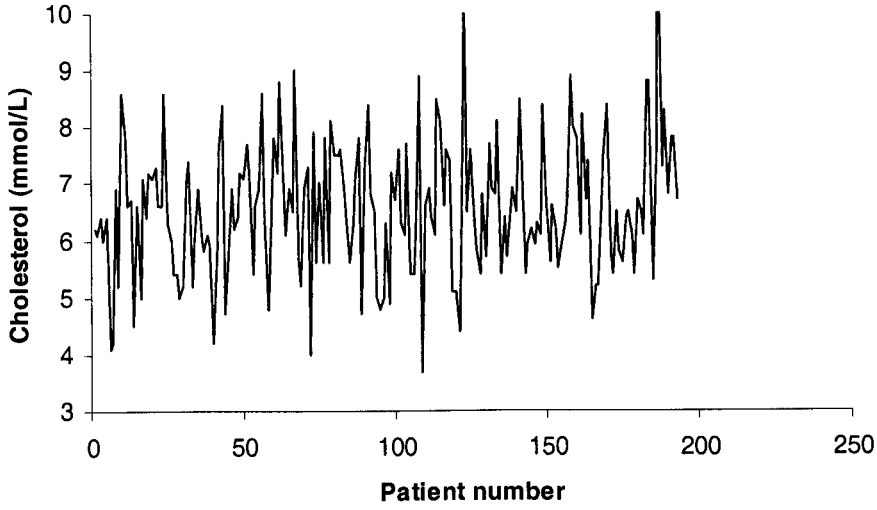
Centre 1



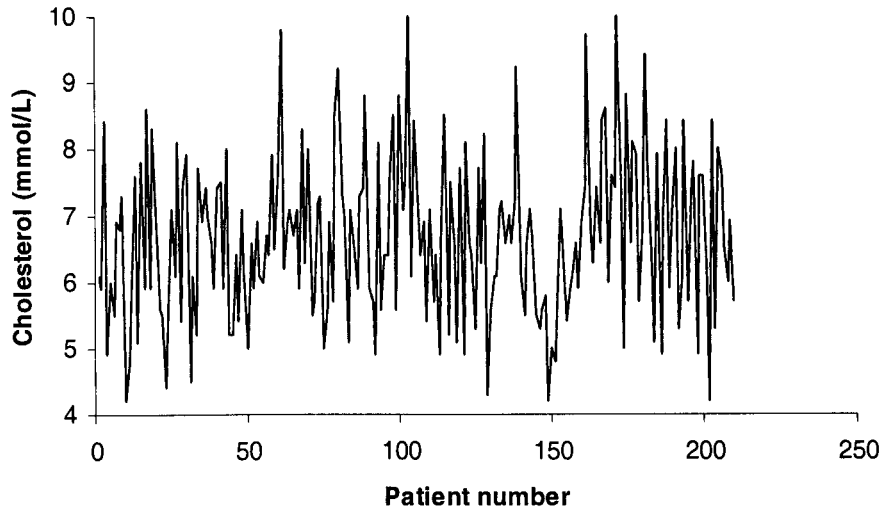
Centre 2



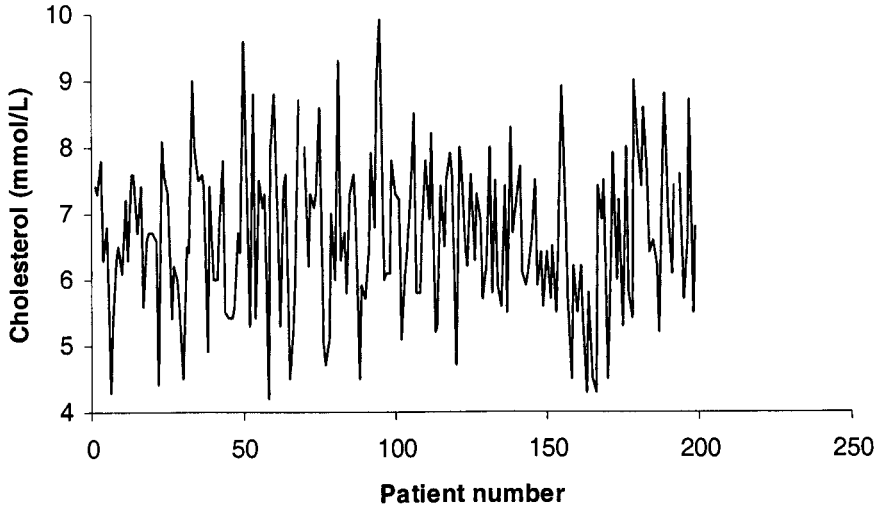
Centre 3



Centre 4



Centre 5



The results from the diet trial show trends over part of the range, that are not present in the drug trial in all centres. There are upward, downward or cyclical trends for the diet trial data. These irregularities are consistent with definite non-random components and are absent from the drug trial.

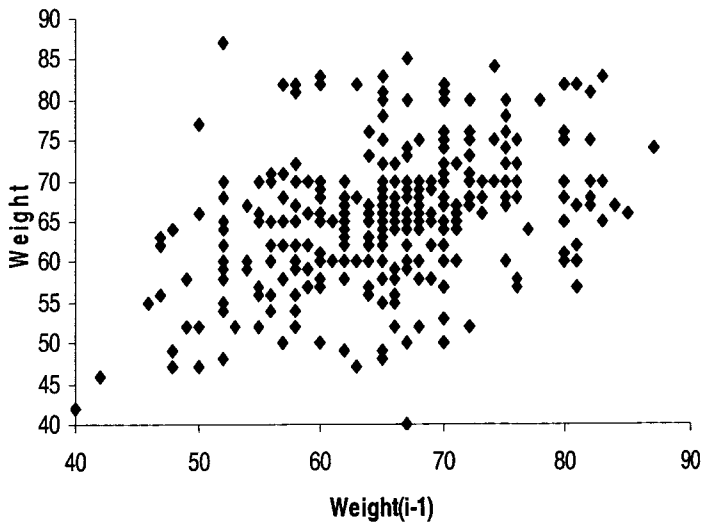
5.4.3.3 The lag plot

The lag plot is a scatter plot of each observation against the previous observation. Random data in terms of relation between successive observations, and inclusion in a trial is generally a random process, should not exhibit any identifiable structure in the lag plot. The lag plots is shown for the weight, systolic blood pressure, diastolic blood pressure, and cholesterol measurements in Figure 5.3a for the intervention and control groups in the diet trial and Figure 5.3b for the same variables for centre 1 in the drug trial. The correlation coefficient and the significance level for all variables and all centres in the diet and drug trials are shown in Table 5.7.

Figure 5.3a Lag plots for the intervention and control groups at baseline in the diet trial

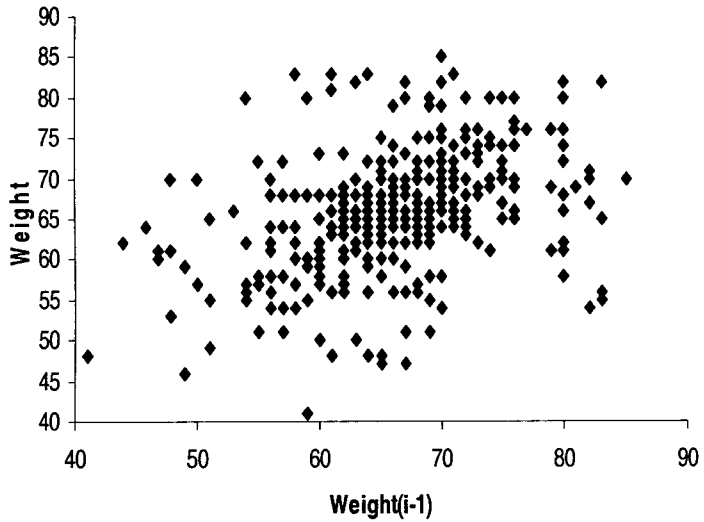
Weight

Intervention group



r= 0.42 ($p=4\times10^{-19}$)

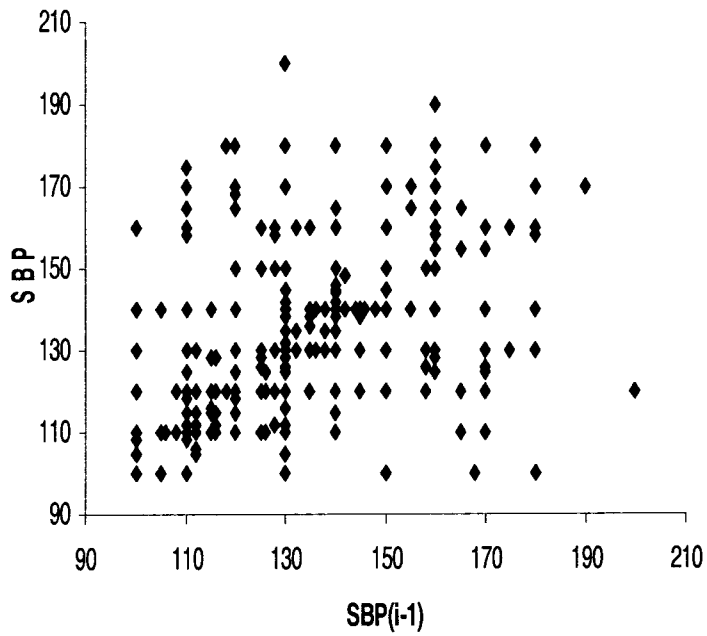
Control group



r=0.453 ($p=2\times10^{-22}$)

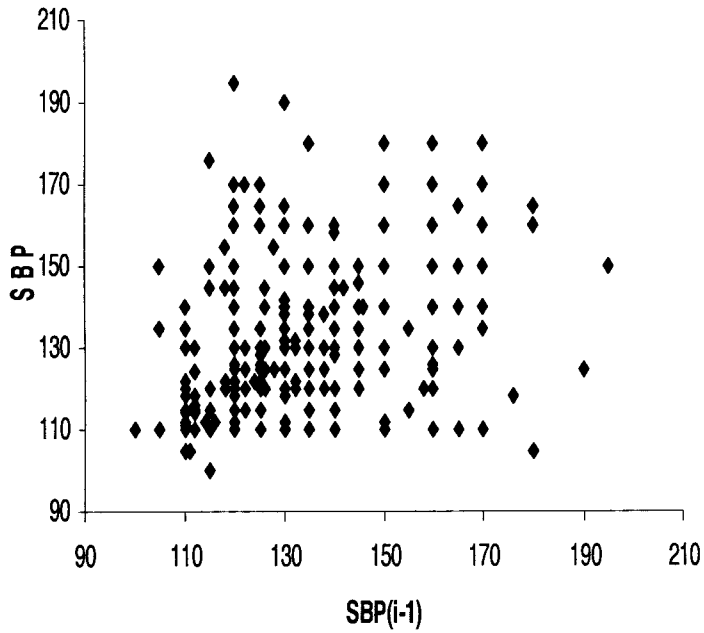
Systolic blood pressure

Intervention group



r= 0.406 ($p=7\times10^{-18}$)

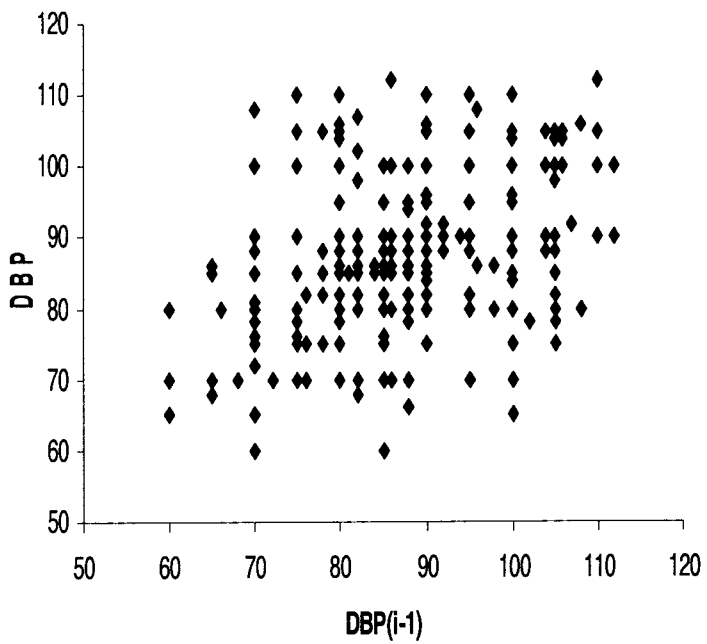
Control group



r=0.41 ($p=3\times10^{-18}$)

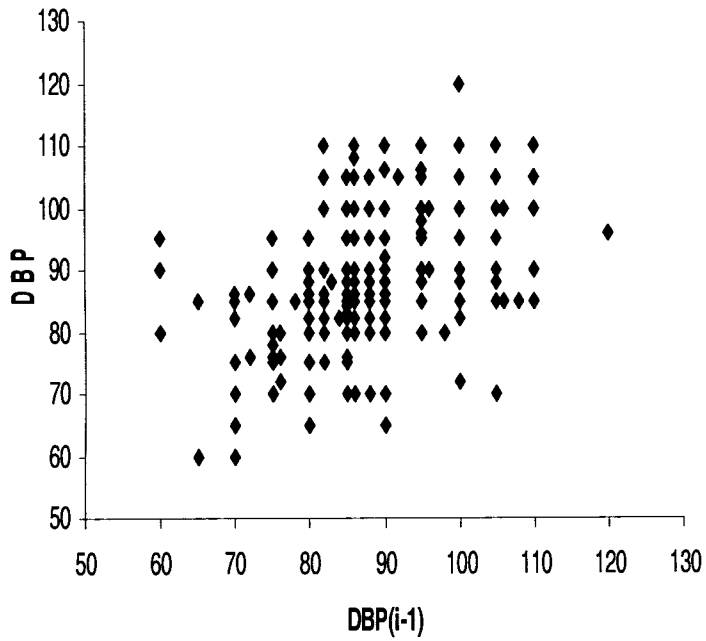
Diastolic blood pressure

Intervention group



r=0.44 ($p=5\times10^{-21}$)

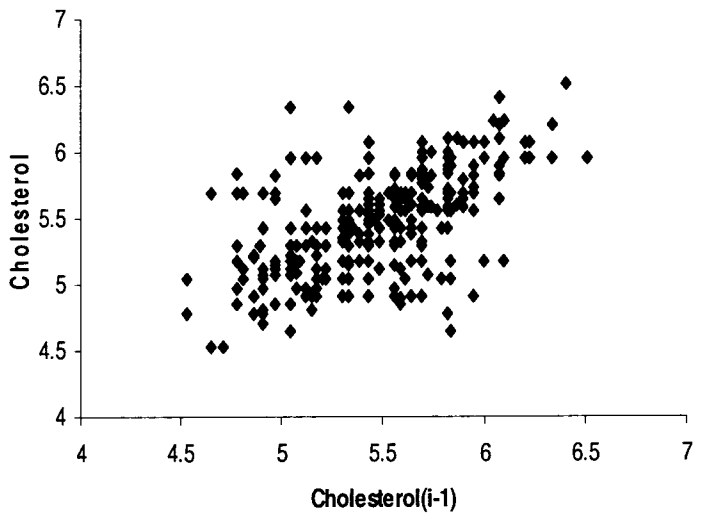
Control group



r=0.459 ($p=5\times10^{-23}$)

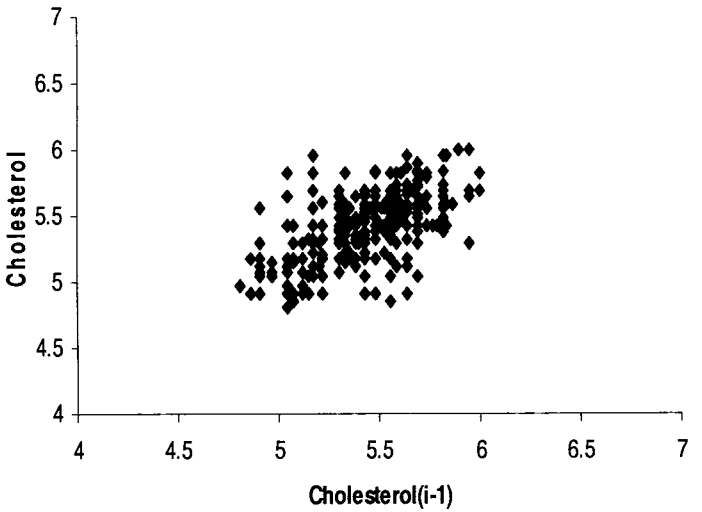
Cholesterol

Intervention group



r=0.616 ($p=1\times10^{-44}$)

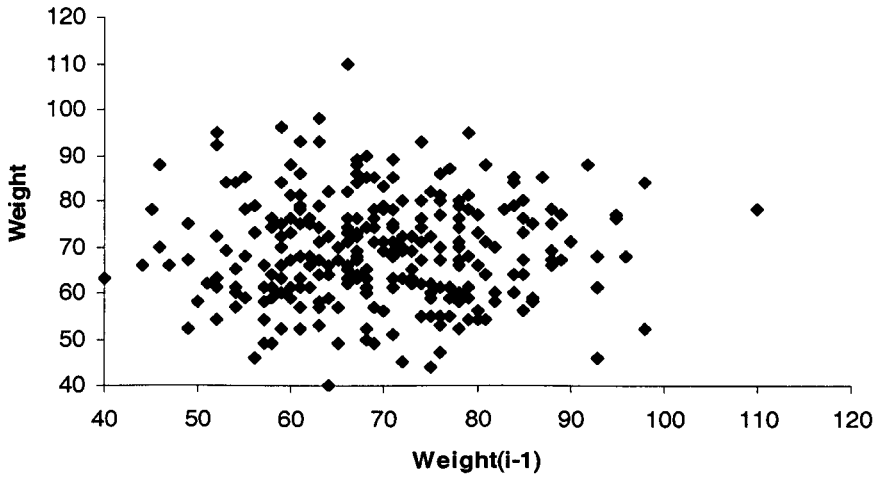
Control group



r=0.362 ($p=3\times10^{-14}$)

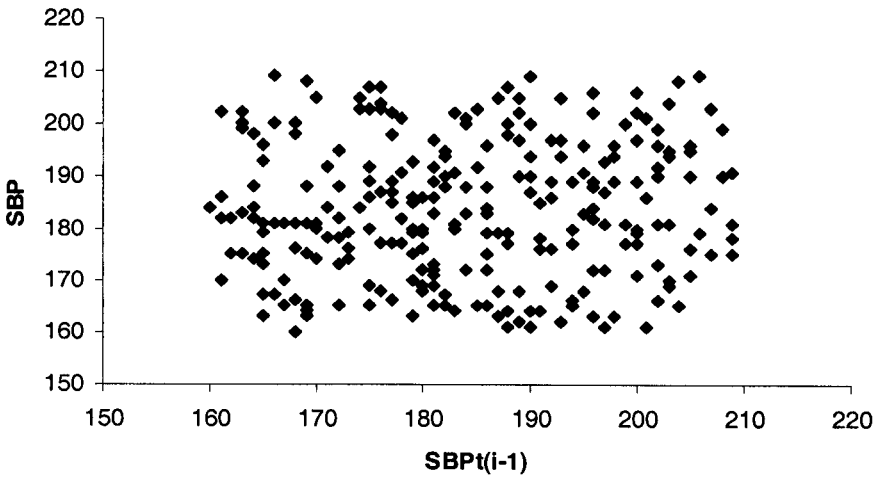
Figure 5.3b Lag plots for centre 1 at baseline in the drug trial

Weight



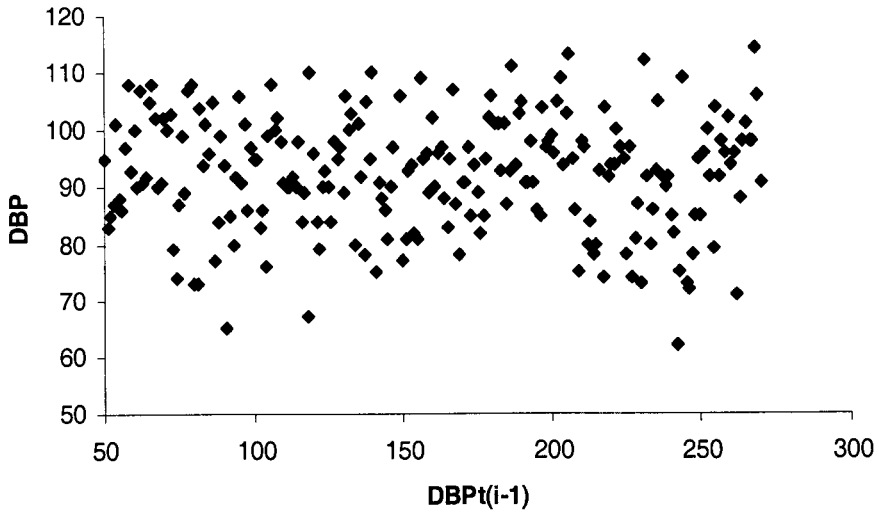
r= 0.011 (p= 0.862)

Systolic blood pressure



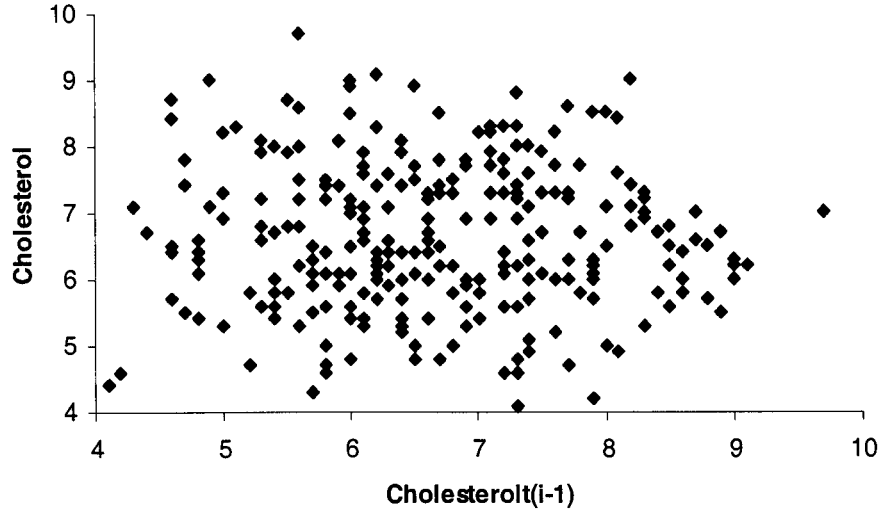
r= 0.07 (P= 0.252)

Diastolic blood pressure



r= 0.152 (P= 0.013)

Cholesterol



r= 0.007 (P= 0.909)

Table 5.7 The correlation coefficient (with p value) between each observation and the previous in the diet and drug trials

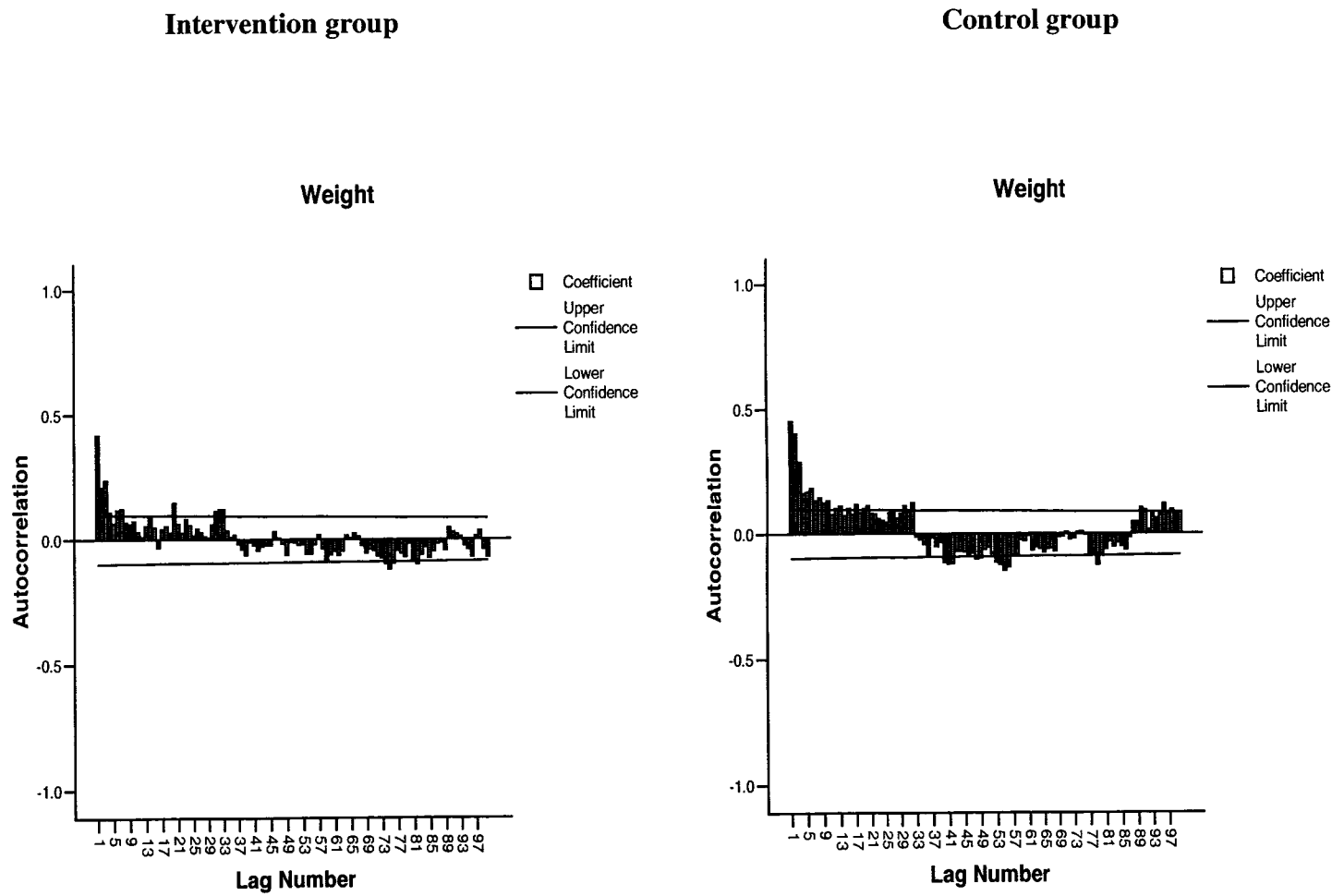
	Diet trial		**Drug trial**				
	Intervention	Control	Centre 1	Centre 2	Centre 3	Centre 4	Centre 5
Height			.2 (.007)	.26 (.001)	.065 (.37)	.217 (.002)	.045 (.53)
Weight	.43 (4×10^{-20})	.45 (1×10^{-23})	.01 (.86)	.095 (.213)	-.054 (.45)	.182 (.008)	.066 (.36)
SBP	.41 (4×10^{-19})	.41 (2×10^{-19})	.07 (.25)	-.032 (.68)	.001 (.99)	-.043 (.53)	.082 (.25)
DBP	.44 (2×10^{-21})	.45 (3×10^{-22})	.15 (.013)	.048 (.53)	-.013 (.86)	.06 (.39)	.047 (.52)
Cholesterol	.61 (8×10^{-44})	.28 (6×10^{-9})	.007 (.9)	-.013 (.87)	.124 (.09)	-.039 (.57)	.069 (.34)
Fasting blood glucose	.25 (2×10^{-7})	.36 (3×10^{-14})					
Total cholesterol	.34 (2×10^{-12})	.44 (2×10^{-21})					
Triglycerides	.19 (7×10^{-5})	.33 (3×10^{-12})					
Energy	.41 (8×10^{-18})	.49 (2×10^{-26})					
Total carbohydrates	.48 (1×10^{-25})	.48 (1×10^{-24})					
Complex	0.6 (7×10^{-42})	0.5 (1×10^{-24})					
Protein	.43 (2×10^{-20})	.42 (9×10^{-19})					
Fat	.47 (1×10^{-23})	.49 (6×10^{-27})					
Saturated fat	.46 (2×10^{-23})	.48 (1×10^{-24})					
Fibre	.53 (3×10^{-31})	.49 (2×10^{-26})					
Soluble fibre	.51 (2×10^{-28})	.67 (7×10^{-55})					
Caffeine	.38 (5×10^{-16})	.61 (3×10^{-44})					
Salt	.56 (7×10^{-36})	.37 (1×10^{-14})					
Vitamin C	.63 (7×10^{-48})	.59 (1×10^{-40})					
Carotene	.62 (5×10^{-45})	.14 (0.005)					
Vitamin E	.58 (4×10^{-38})	.56 (4×10^{-35})					
Vitamin A	.94 (3×10^{-193})	.37 (4×10^{-15})					

Non-random structures in the lag plots indicate that the underlying data show first order autocorrelation, where each value is correlated with the previous participant's value. These serial connections in the values should not occur in genuine data (where the case order should be effectively random). Figure 5.2a shows distinct positive autocorrelation in the diet trial, a good indication of non-randomness. While one or two variables might show some serial correlation, as with the drug trial, what is unusual is that so many of the variables show this effect, and the probability that the different variables show such an effect when they should be independent is so incredibly small that it could not have arisen by chance.

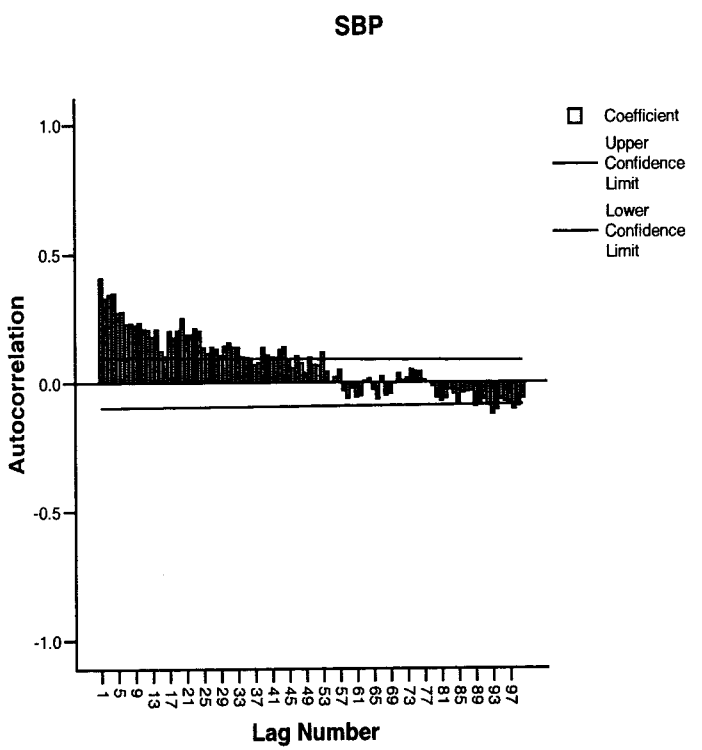
5.4.3.4 The autocorrelation plot

Autocorrelation occurs when an observation is, in some part, determined by the preceding observations. This is a very common kind of non-randomness and extends the results from the runs test and Figure 5.3 to all spacing between values, called lags. If the lag is one, the first observation is compared to the second, the second with the third and so on; if the lag is two the first observation is compared to the third, the second with the fourth and so on. The autocorrelation plot below for the first hundred lags (Figure 5.4a) suggests that the data in diet trial are not random. The limits in these plots, denoted by horizontal lines, give a 95% range within which the calculated correlation coefficients should lie, given no underlying autocorrelation. For diet trial, most of the values are out of control (outside the limit) with many of the autocorrelations too large. This indicates that the data do not hold the property of independence, as there is much dependence between successive observations for this trial. No such discrepancy is seen for the drug trial, for instance, centre 1 Figure 5.4b.

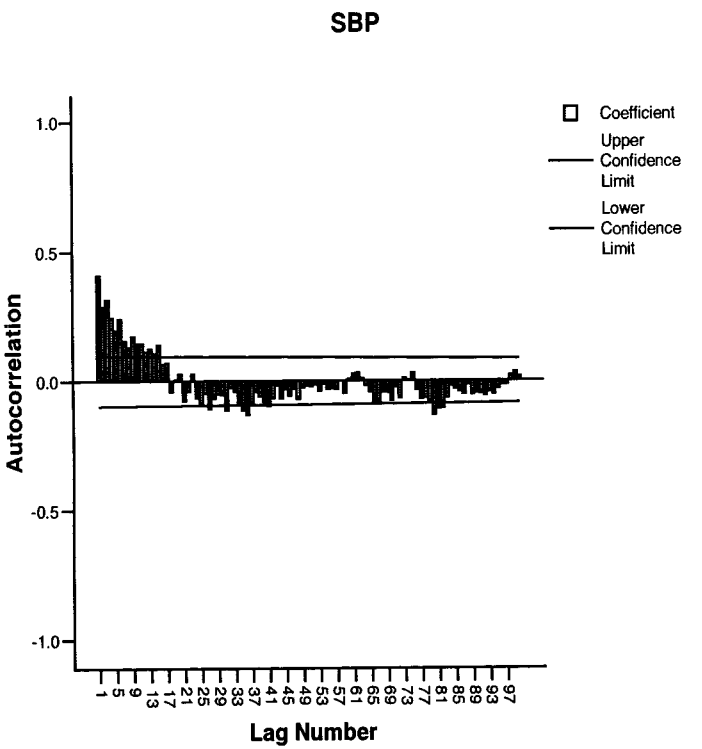
Figure 5.4a Autocorrelation plots for the intervention and control group at baseline in the diet trial for the first 100 lags



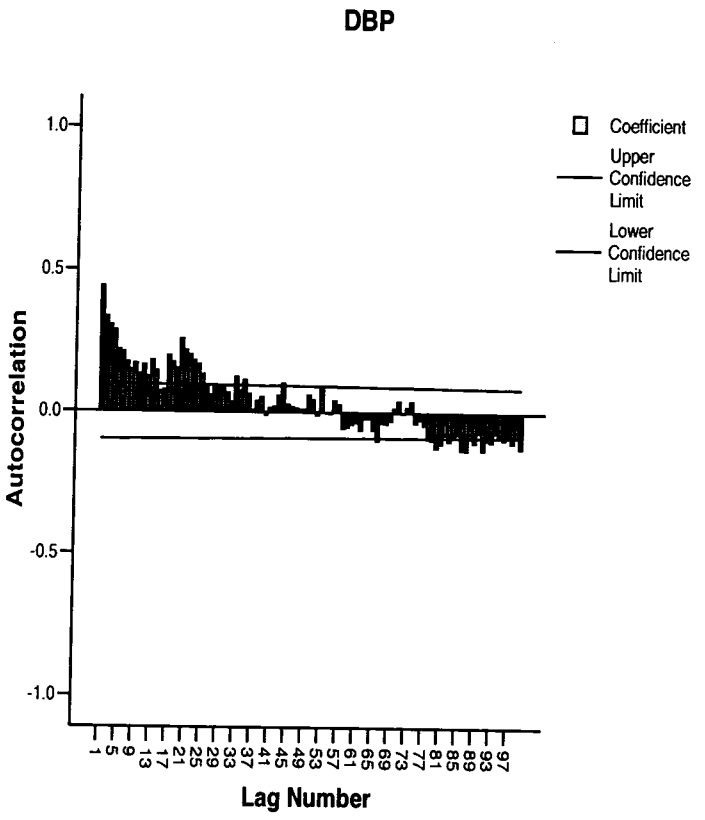
Intervention group



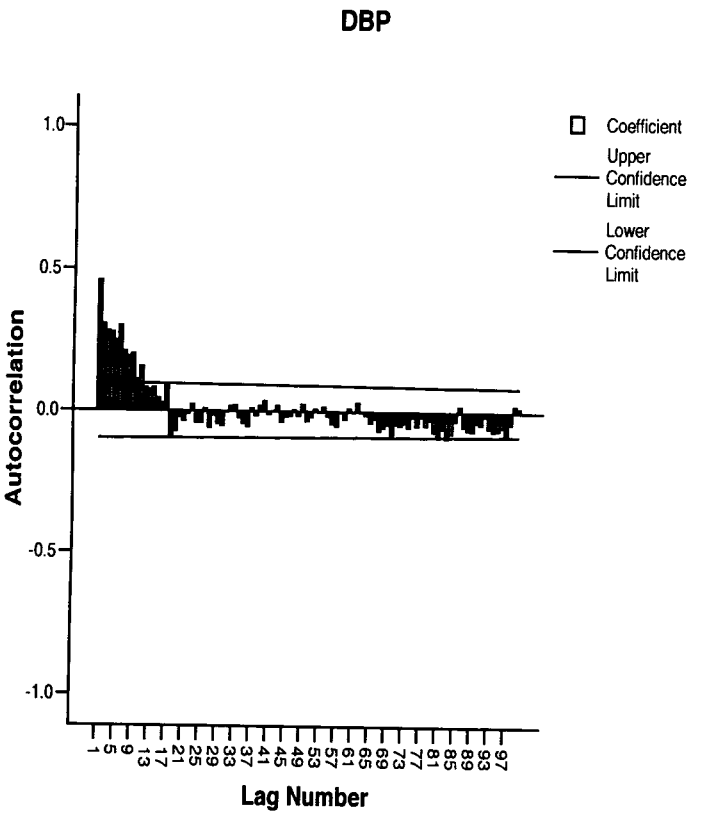
Control group



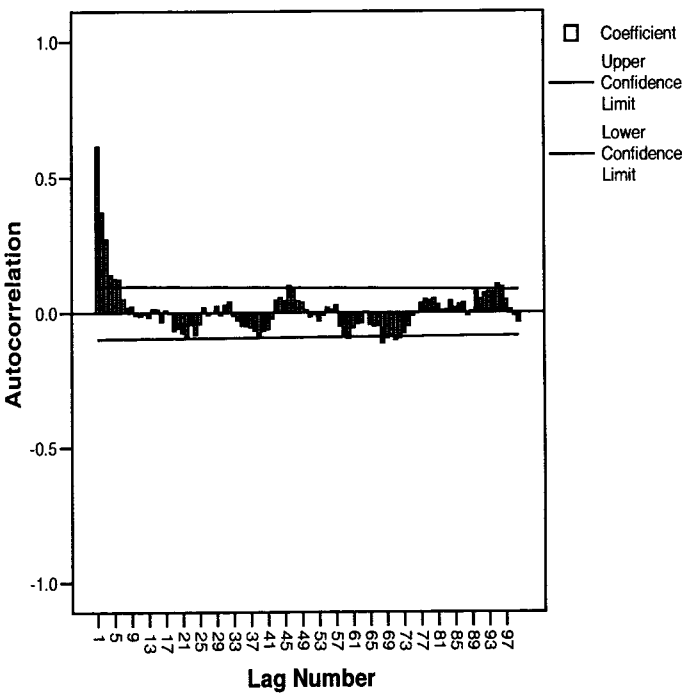
Intervention group



Control group



Intervention group



Control group

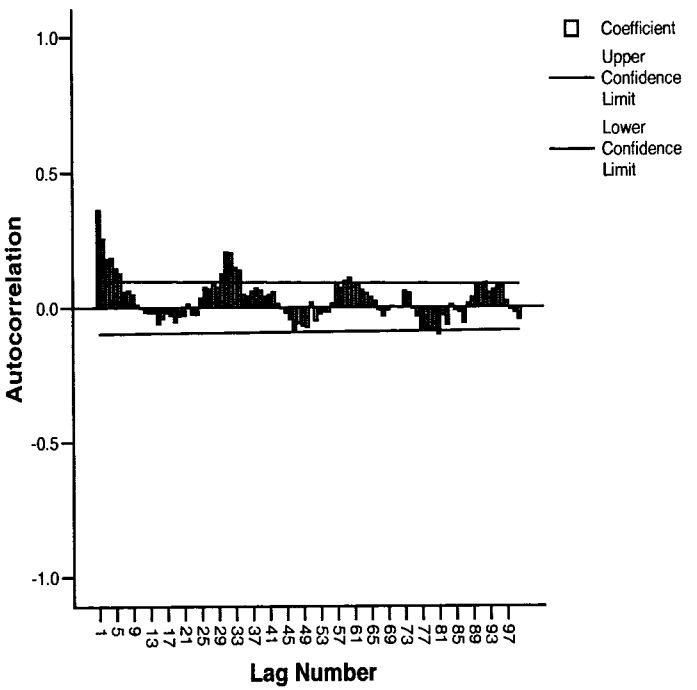
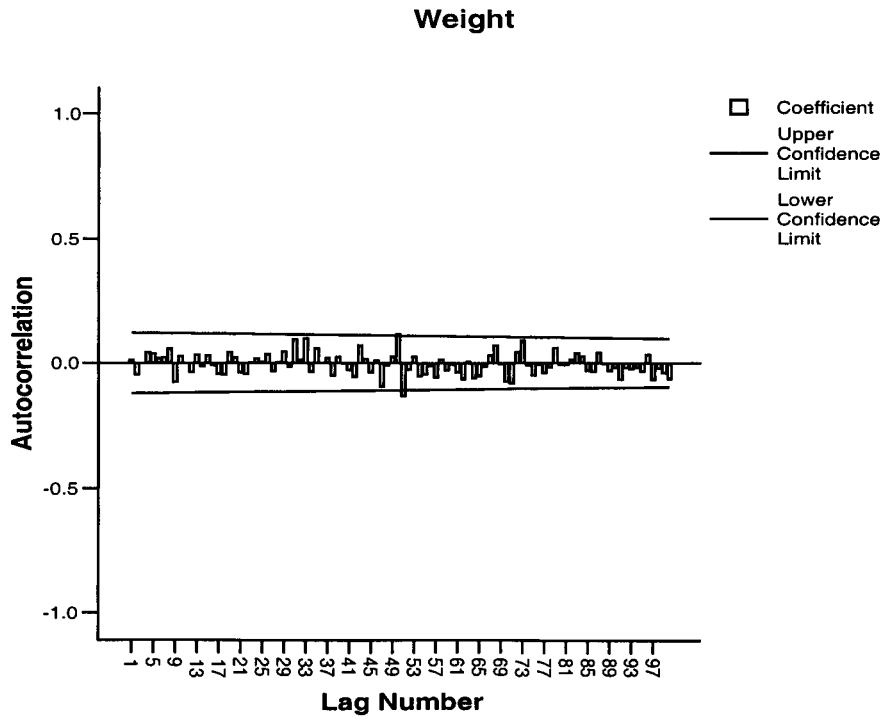
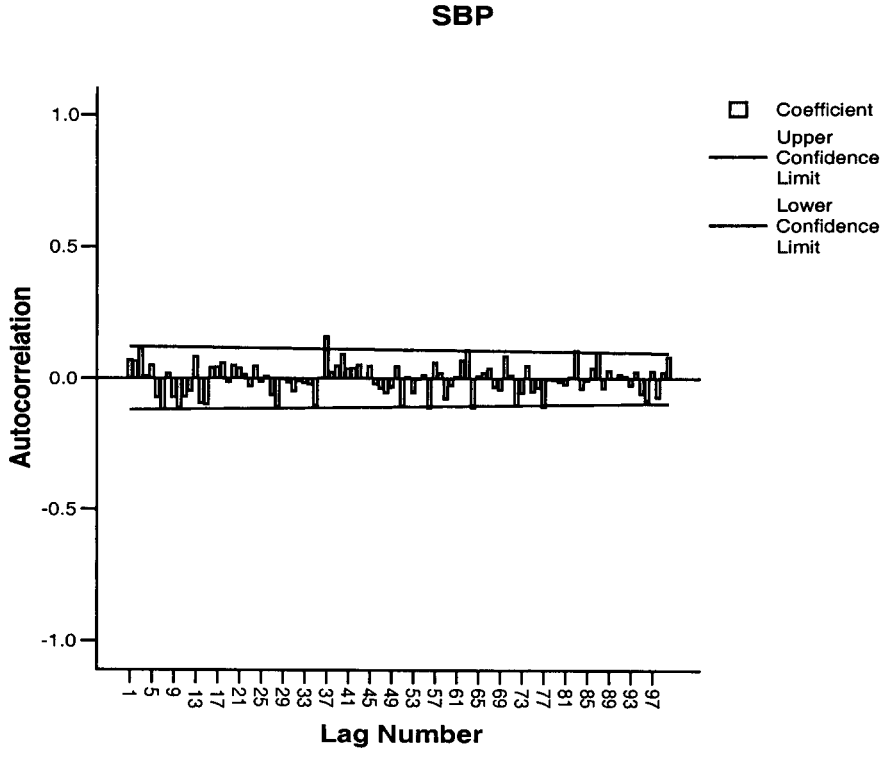
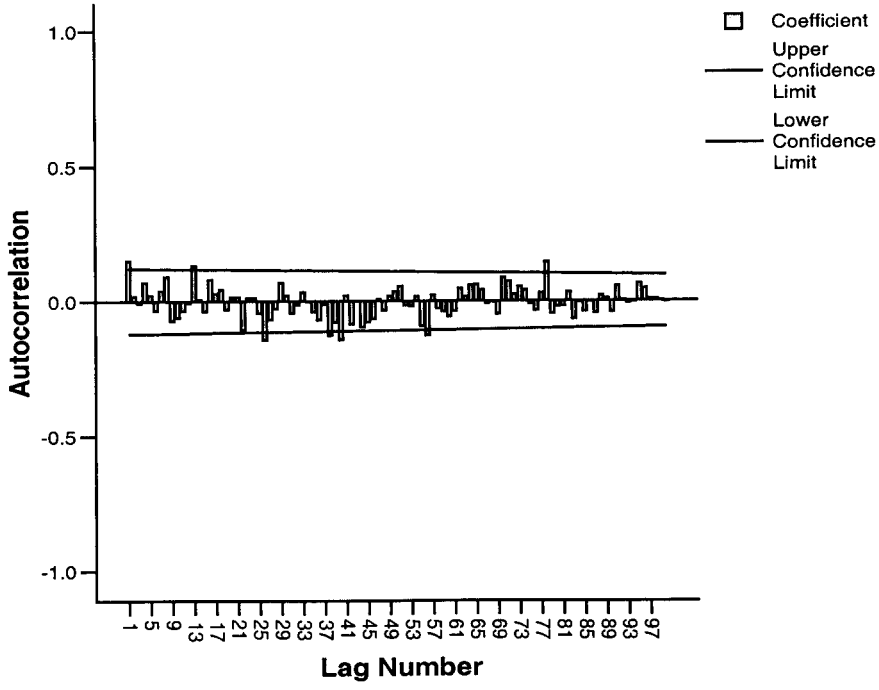


Figure 5.4b Autocorrelation plots for centre 1 at baseline in the drug trial for the first 100 lags

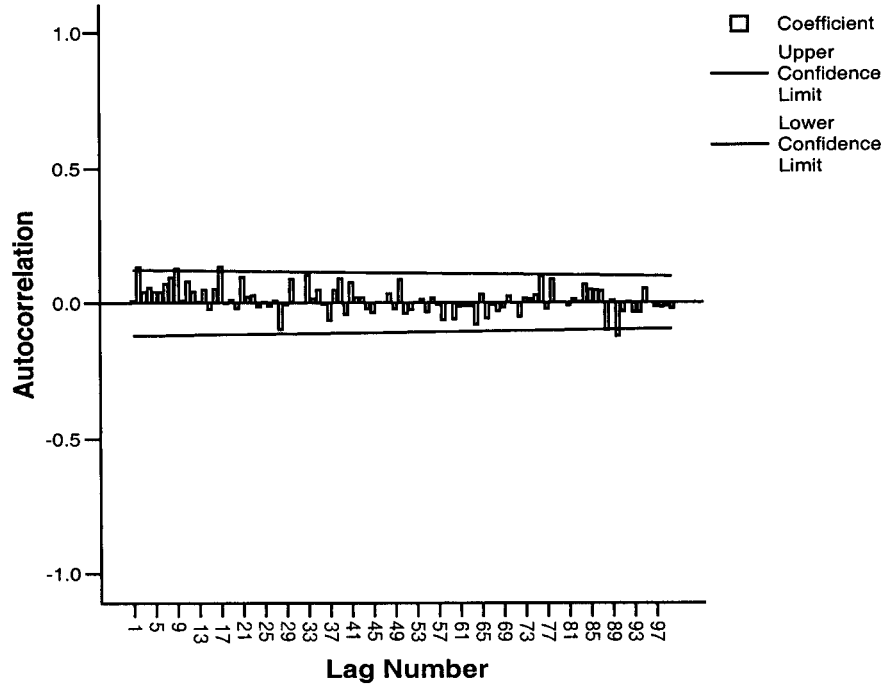




DBP



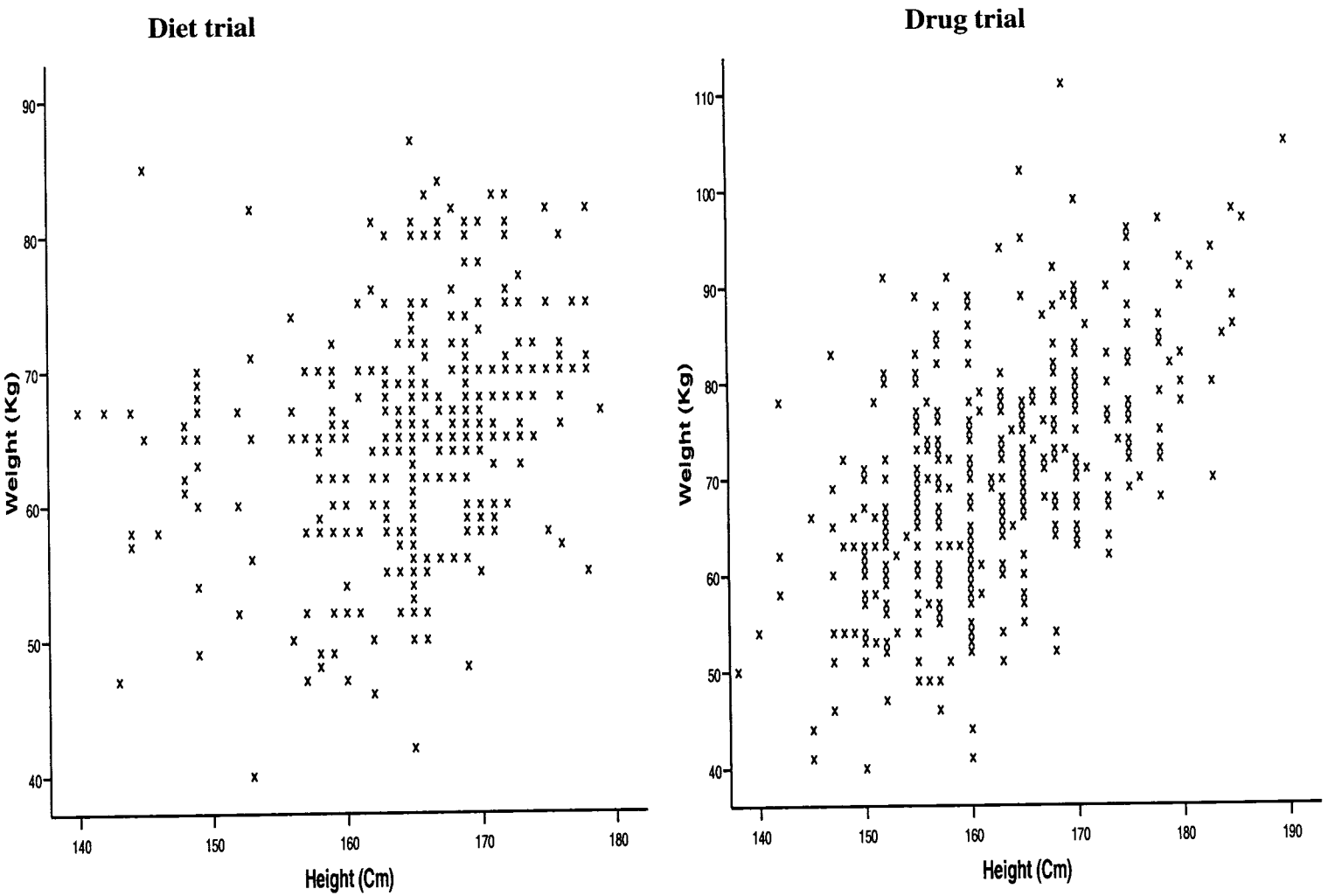
Cholesterol



5.4.3.5 The scatter plot

The scatter plot is a plot of the values of one variable versus the corresponding values of the other one, with one point per subject. It reveals the relationship or association between two variables. Here, the relationship between height and weight are explored in the two trials. In general, height and weight are related; with taller people tending to be heavier than shorter people. The relationship is not perfect; people of the same height vary in weight. The scatter plots below (Figure 5.5) shows the relationships between height and weight values in the diet and drug trials. The pattern clearly shows a positive relationship, height and weight tend to go up and down together in the two trials; however, the correlation coefficient in the drug trial is stronger and this stronger relationship between the two variables in the drug trial is clearly seen in the diagram. A linear relationship can be seen in both but it is stronger and much more definite in the drug trial; in the diet trial, the relationship is patchy and weaker for lower heights.

Figure 5.5 Scatter plot for the intervention groups at baseline in the diet and drug trials



r= .301

r=.557

5.5 Discussion

Every examination of the data from the diet trial is consistent with some form of misconduct. Parallel examination of data from the drug trial does not show any hint of misconduct, or unusual or unexpected features that would lead to a suspicion of misconduct.

One strength of the comparison is that the drug trial was multi-centre and for some of the tests, five of the centres were compared separately with the diet trial. A potential weakness was that, the admission order number could not be obtained for each patient in the drug trial, thus it has been assumed that the patients' enrolment to this study was as the date of randomisation.

The first absolutely definite conclusion drawn from the fact that the means and variances (standard deviations) in the diet trial are so different between the groups for so many variables at baseline which means that the trial cannot have been randomised in any normal sense of the word. It is not a randomised controlled trial (RCT). There is no explanation of the pattern seen that is compatible with a true RCT.

The fact that so many different variables have such high correlations between successive observations makes it extremely unlikely that the data arise from a genuine study and they add to the suspicion of misconduct. The difference in digit preference between the groups adds strongly to the evidence also that this is not an RCT. If this is not an RCT, the question is how the data arose.

A possibility is that the randomization processes was subverted, but the data themselves were genuine. If this was so, one expects consistency in the mean and variance between the groups. Here highly significant differences in variance are found, but none in the mean for height, SBP, DBP, cholesterol, fasting blood glucose, triglycerides, and salt but

highly significant differences in both (variance and mean) for complex, protein, fat, saturated fat, fibre, soluble fibre, carotene, vitamin E, and vitamin A. There is also a significant difference in the mean but not in the variance for vitamin C. More consistent differences would have been expected if there were a tendency to put for example; higher blood pressures into one group, at least among variables correlated with blood pressure.

However, the different patterns of digit preference are not compatible with subversion of the randomization process to put patients at say, high risk, into one group but not the other alone. These patterns of digit preference could be compatible with one person recording the data for one treatment group and another person recording the data for the other group. If the allocation of cases was not concealed from the person reading data at the time of data collection, and the two groups were recorded separately, and if only the digit preference differed, this might be an explanation; the person recording the data could be different for each group and then they applied their own rounding idiosyncrasies to the results as written down. Thus, one explanation to suggest that randomization was simply subverted to alter means is incompatible with another explanation to suggest that innocent errors in rounding occurred differentially between groups.

It has been shown that statistical techniques work and show fraud in this dataset. Therefore, it could be used for other data on suspicion of fraud. It is important to remember that these statistical techniques may not work if the fraudster uses computer based methods for making up data and with technical expertise.

Reference:

1 Armitage P, Berry G. *Statistical methods in medical research.* 3rd ed. Oxford: Blackwell Scientific, 1994:386-401.

2 Altman DG. *Practical statistics for medical research.* London: Chapman and Hall, 1991:122-143.

3 Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B, et al. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Stat Med* 1999;18:3435-51.

4 Taylor RN, McEntegart DJ, Stillman EC. Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. *Drug Inform J* 2002;36:115-25.

5 White C. Suspected research fraud: difficulties of getting at the truth. *BMJ* 2005; 331:281-288.

6 Medical Research Council Working Party. MRC trial of treatment of mild hypertension: principal result. *BMJ* 1985;291:97-104.

7 Evans S. Statistical aspects of the detection of fraud. In: Lock S, Wells F, Farthing M, eds. *Fraud and misconduct in medical research.* 3rd ed. London: *BMJ* Publishing Group, 2001:186-204.

CHAPTER 6

Discussion

6.1 Overview

A consensus statement developed at a UK Consensus Conference on Misconduct in Biomedical Research, organised by the Royal College of Physicians of Edinburgh, defined research misconduct as "behaviour by a researcher, intentional or not, that falls short of good ethical and scientific standards" (1). The definition of research misconduct should not be restricted to fabrication or falsification of data. It should cover the whole range of research misconduct.

A high standard of medical research practice, which includes integrity, honour, and truthfulness, is essential if public confidence is to be guaranteed. Misconduct is a potential problem in medical research, leading to the possibility of patients being given useless or ineffective treatment, or even denied effective treatment. Any case of misconduct, fraud or corruption of scientific records reduces public trust, leads to false conclusions and causes frustration to careful and honest workers

A preliminary search of PubMed database was carried out to explore the types of misconduct found in clinical trials and reported in the scientific literature from January 2000 to July 2003. 'Clinical trials' AND 'scientific misconduct' were used as keywords. The data were broken down into types of scientific misconduct, showing that data fabrication and falsification were two of the most reported findings.

This small sample of papers provided insufficient information of general frequency or impact of the misconduct reported. The papers also tended to focus on specific instances

of misconduct and a more general discussion of types of misconduct, their importance and relative frequency was lacking. In order to assess the types of misconduct likely to distort the results of clinical trials and likely to occur, it was therefore decided to conduct a Delphi survey, soliciting opinions from experts in clinical trials through questionnaires.

I considered the most important types of misconduct to be those believed to distort trial results and occur most commonly. My results show that >50% of experts suggested that the opportunistic use of the play of chance principally (inappropriate sub-group analyses) and the selective reporting of outcomes are the most important types of misconduct. The results from the Delphi survey were considered as a determinant of the direction of the research reported in later chapters.

Any action to prevent the occurrence of such types of misconduct should address the issues of selective reporting of the results of clinical trials and the opportunistic use of the play of chance. All primary and secondary outcomes should be pre-defined clearly in the protocol and reported completely. Full details of the plan of sub-group analyses need to be justified and written up in the protocols. Any sub-group analyses reported without pre-specification in the protocol would need supporting evidence within the publication to justify them.

To examine the extent and nature of selective reporting and how sub-group analyses were described in protocols of clinical trials, and then how they were reported, forty-eight protocols of randomised controlled trials were reviewed using information available on *The Lancet* website in July 2004. Additionally, publications of thirty protocols were reviewed alongside the protocols.

A clear definition of the primary and secondary outcomes in the protocols was an important problem encountered. Selective reporting of primary and secondary outcomes

and subgroup analyses in the publication of the thirty trials was also very common. In 17%, 73% and 50% of the trials, at least one primary, secondary or subgroup analysis was unreported. Most of the randomised controlled trials' protocols provided unclear information on subgroup analyses. There was noticeable deviation from the protocols in most of the corresponding publications. More primary outcomes appeared in the reports compared to the protocols (60 versus 51) and less secondary outcomes (93 versus 133).

Although the experts in clinical trial methodology involved in the Delphi survey believed that fabrication and falsification of data (fraud) are uncommon, there was consensus that, if it does occur, the findings are highly likely to be distorted. As I mentioned in chapter 1, a result from a survey (2) showed that 0.3% of US scientists funded by NIH, confessed that they had falsified or 'cooked' research data and almost every seventh (15.3%) indicated that they dropped observations or data points from analysis based on a gut feeling. Therefore, there are no guarantees that fraud will not occur and there is a case for biostatisticians to investigate methods of fraud detection on a more routine basis. A systematic review of statistical techniques for fraud detection was therefore conducted.

A number of techniques emerged from the review, some of which were applied in the context of fraud detection. However, several techniques were presented in the literature without examples or references to how they could be applied in practice.

Analytical and graphical techniques were then applied to data from two randomised controlled trials to demonstrate fabrication of data, one of which had already raised suspicion among reviewers, and every examination of data from it was consistent with some form of misconduct. In contrast, when the same techniques were applied to the other data, there was no hint of any unusual or unexpected features.

6.2 Selective reporting versus fraud

Selective reporting was considered by the experts in the Delphi survey to be potentially distorting of the trial results and a very common issue. The experts' view was confirmed by the finding from reviewing the Lancet protocols and corresponding publications. It is rare to find reports from clinical trials which do not contain selective reporting of findings. Such practice seems to be universally accepted by investigators, editors and readers and sometimes even considered within the science culture's nature (3). In contrast, deliberate fabrication or falsification of data would never be acceptable and no one would condone a study where fraud could happen as a part of its process.

The exact prevalence of selective reporting in clinical trials is unknown. Many authors try to highlight positive findings and downplay negative outcomes, in order to make their manuscripts more interesting and increase their chance of publication. Under-reporting of non-significant findings make the significant ones seem superior, which introduces a distortion in the final picture. One may expect around one in twenty comparisons to be statistically significant at 5% level by chance alone. If the significant one is the only one reported and the other 19 (non-significant) are not, the reader is misled.

Not reporting all the outcomes also has potential consequences for the patients and wastes of limited resources. When an ineffective intervention is falsely reported as effective (by chance), it can lead to patients receiving the ineffective treatment and being denied an effective one.

Overestimating the efficacy of the new intervention causes a problem; as the new interventions are usually more expensive than the conventional once, this would increase the cost of therapy without corresponding enhancement in the outcome.

Fraud (data fabrication and data falsification) perhaps is the extreme type of scientific misconduct. It is a severe public offence, but probably an uncommon one. The real problem is that if all or most of the data of a research project have been fabricated or falsified which would affect the conclusion from that research. Then the literature could be build upon what has been falsely reported. The problem in clinical trials is worsened because of the potential harm to patients, who may be recommended treatments incorrectly.

If a scientist is found to have committed fraud in one study, all the data in his /her work may be fraudulent. Researchers and systematic reviewers need to know about all that scientist's work.

The diet trial data by Dr Ram B Singh in Chapter 5, which was confirmed to be fraudulent is similar to the one used for the analysis and published in the *Lancet* in 2002 (4) by the same author. This study was a randomised controlled trial on the effects of an Indo-Mediterranean diet on the progression of coronary artery disease in high-risk patients. One of its findings was that an Indo-Mediterranean diet rich in a-linolenic acid might be more effective in primary and secondary prevention of coronary artery disease than the conventional prudent diet. Since the publication of this trial, much has been written on the benefits of the Mediterranean diets. This study has been cited 26 times, including in guidelines, and its lead author went on to publish many papers in other journals. Regardless of the final decision on the paper, there is doubt about its claim regarding the potential benefit of the dietary measures was tested.

In multi-centre clinical trials, if the data in one centre with few patients are fraudulent, this is not of extreme concern because of the small number, so these data would be unlikely to affect the conclusion.

The relative magnitude of the effect of selective reporting or fraud in clinical trials depends on the number of studies where fraud has taken place and the number of studies where there has been selective reporting of outcomes or sub-groups analysis, without adequate description and appropriate statistical analysis.

I conclude that both types of misconduct, selective reporting and fraud, can undermine the conclusion of a trial and its impact in the literature. Persons who carry out fraud strongly hope that the fraud goes unnoticed. On the other hand, selective reporters in general have no hesitation in practicing this or in hiding what they do. A general level of tolerance of selective reporting leads to it occurring frequently and generally being scientifically accepted. In my opinion, selective reporting has introduced a greater problem than fraud due to its frequency of occurrence. Hence, I believe that a specific system should be adopted to control selective reporting in publication of clinical trials.

6.3 Some proposed issues to control inappropriate sub-group analyses and selective reporting in clinical trials

The placing of study protocols in the public domain has been discussed (5,6-8). This might help to deal with issues of discrepancy between protocols and published reports. Comparing what was originally planned and what was actually done at the time of a journal review of a submitted report, to check the deviation would then be possible for any independent investigator. The findings from this study supported the findings from the previous ones (5,9), which demonstrated major discrepancies between trial protocols and subsequent publications. They also showed that selective reporting of outcomes in published randomized controlled trials is common. It is worth noting that this occurred even with the protocols listed in *the Lancet*. However, the full protocols were not in the public domain, just brief summaries of accepted protocols were on the web and authors'

consent was needed to access them. Placing only protocol summaries in the public domain is not a sufficient solution to the problem.

A stipulation that authors must submit trial protocols at the time of manuscript submission does not completely address the problem of selective reporting and inappropriate sub-group analysis, since authors could alter the protocol before this submission.

Guidelines and recommendations for the design, analysis, interpretation and reporting of sub-group analyses have been proposed (10,11). However, it seems that these guidelines are not being implemented and many researchers fail to appreciate them and fail to report their studies adequately.

A new system has been proposed (12) to improve the quantity and quality of reporting of clinical trial results. This proposes that a systematic review of the evidence supporting the need for the trial is prepared and placed on the Web. After registration of the trial, the full trial protocol based on the conclusion of the review would also be posted on the Web. The statistical analysis plan in the protocol would be pre-specified and pre-programmed. When data collection is completed, the dataset would be uploaded and the analyses run. The results are placed in the public domain (on the Web) for comments. Results from this analysis will be added to the systematic review. Anybody can comment online on the trial design and contribute toward appropriate analysis of the results. Journal publications would be concerned commentary on the systematic review which now includes the trial results.

This system could eliminate bias of individual trial results, as only the pre-specified analyses are run. However, it would expose the idea and the design of the trial before it is finished, which might raise problems of confidentiality. Currently assessment of the

publication of clinical trials requires peer reviewers, who are not involved in this system. More groundwork is needed before this proposal can be seriously implemented.

Despite these suggestions, selective reporting of research results and inappropriate sub-group analysis still exist. This problem would be considerably reduced if primary and secondary outcomes and sub-group analysis are defined clearly in the protocol (which was not the case in most of *the Lancet* protocols), if all findings are reported in full, and if subgroup analysis is carried out adequately. My suggestion to achieve this starts making improvements in the quality of the protocols of randomized controlled trials, by posing full protocols on the web as described below.

Electronic submission of protocols of randomised clinical trials

With these considerations in mind, I suggest that the procedure of publishing clinical trial results should take place in three primary steps.

The first is trial registration. The second is the electronic submission of the clinical trial protocol, and the third and last step is the electronic submission of results.

The International Committee of Medical Journal Editors (ICMJE) has proposed comprehensive registration of clinical trials at the time of their conception (13). Registration of trials can help anyone with an interest in an area of medical practice find information about every clinical trial in this area. A minimum registration information data set includes 20 items. Registration information will be considered inadequate if it has missing items or items that contain uninformative terminology.

The second step is the most important. Web-based submission of the protocols of clinical trials has been suggested (14) to overcome the space constraints in printed journals. However, a more efficient and detailed web-based submission of protocols could be very

important to ensure reliability and consistency in conducting the trials throughout. Prominent journals should adapt a standard form for submitting a full web-based protocol and post it on their website. Investigators would decide the journal in which they want to submit their protocol and then to publish the manuscript of the results from that study. A further restriction is that the submission of protocols must be online only. Neither protocols submitted by fax nor by post would be accepted. The online submission of protocols would oblige the investigator to provide a complete protocol.

An accepted protocol must contain all information about the trial. All fields of the standardized form should be completed by the investigator. The system would reject the submission of incomplete forms, and request the completion of missing data in fields. Moreover, when entering the number of primary and secondary outcomes, and the number of subgroup analyses, the system would generate a fixed number of input fields, equal to the number of outcomes or subgroups provided by the investigator earlier in the form. For example, if the investigator reported 3 secondary outcomes, the system would automatically create 3 input fields to be filled in with the associated outcome. By this strategy, the investigator would be committed to the exact number of primary, secondary outcomes and sub-group analyses. Once the form is completed, the system generates dummy table for the results to be filled in after the trial is completed. The protocol would be reviewed by the journal and then stored in a confidential and secure third-party repository.

It would be possible to publish a summary of the protocol while the trial is being conducted. The investigator would be allowed to modify the protocol within a set time before the data collection. When this time has expired, no modification would be allowed.

The third and final step of the publishing process takes place after the trial is completed, is results submission. At this stage, the protocol is released on the Web, and is used in the reviewing process for the paper report.

Such a regime could be enforced by a requirement among prominent journals and/or by regulatory agencies (for drug trials), who are able to refuse publication or licensing to non-compliers. General research about the feasibility of these innovations is needed. The proposal would only work if there is strong consensus among journal editors and reviewers. Such a radical change may need further evidence on its acceptability and practicality to support it.

6.4 Strengths and weaknesses

Different types of scientific misconduct in medical research are discussed in the literature. Unfortunately, the full extent of the problem of scientific misconduct in medical research is not easy to identify. After all, perpetrators of fraud will usually try to conceal their activities. Even so, it is difficult either to set out comprehensive and precise definitions of scientific misconduct or details about the prevalence of misconduct, and its impact on the scientific literature. Thus, much discussion to assess this problem has taken place (15-17), but it is fair to say that a focussed consensus has not really emerged. This research considered what types of misconduct were stated by the experts in clinical trials as being important (Delphi survey).

There are some strengths and limitations of the use of the Delphi technique. Delphi has three important strengths. The first is anonymity and separation: the Delphi technique brings out participants' opinions without bringing them physically together. It also avoids the expense and inconvenience of travelling to and from meetings. Thus, it reduces the effect of dominant individuals and allows the group to share responsibility.

Secondly, it allows individuals with the appropriate knowledge in the content area to have different perspectives and points of view, thus providing a flexible way for individuals to approach difficult questions and solve complex problems (18). Thirdly, the statistical analysis of group responses ensures that each expert's opinion contributes to the final response that eventually provides the outcome. These three features of the Delphi technique make it a useful procedure for developing consensus and finding out whether there is enough common ground for claiming the generality of the consensus.

The Delphi also has three limitations. The first is a lack of agreement regarding the size of the panel. However, the use of participants who have knowledge and an interest in the topic may help to increase content validity. Secondly, there is no evidence that results obtained though this technique may be reproducible, or how often non-response (round by round) may occur during the study before it jeopardises the validity of results. The third limitation of the Delphi technique is that the meaning of consensus is undecided. A universally agreed consensus mark does not exist, and indeed is often cited as a major deficiency in studies using the technique (19). However, the final round will usually show convergence of opinion (20). Consensus levels can fluctuate between 51% and 80% (21). Other authors suggest that the stability of the response through the different rounds is a more reliable indicator of consensus than a percentage agreement at the end (22).

Depth and more strength were added in the review of the protocols accepted by *the Lancet* and comparisons between the protocols and the published papers. By comparing the protocols with the published reports, a deficiency was demonstrated, firstly in the definition of primary and secondary outcomes and sub-group analyses (in which there was a lack of information) and subsequently, in that most of the published reports had inadequate reporting of such outcomes. The inadequate specification of sub-group

analysis in many protocols and also the mismatch between protocols and final reports shows that the proposition that protocol review will reduce the problem of inappropriate sub-group analysis needs to be explored further.

The final part of this research demonstrates how statistical techniques can be applied to detect fraudulent data, when patients' records are available. However, it would be difficult to detect data fabricated using computers to generate the data under a sufficiently specified model. This implies that detection of fraud should include inspection of sampled individual patient records or even access to the patients themselves. This is unlikely to be a realistic option, but it could be an argument for authors being asked to state that a minimum percentage of raw data and patient files have been assessed, for quality assurance of the data.

The comparison between randomised groups at baseline, which in general does not strictly need be conducted in a randomised controlled trial, can reveal unexpected features in the data. Our research carried out into the diet trials yielded some unexpected results. Therefore, it is pertinent to report, either that the tests on the baseline had been carried out and were not significant (thereby confirming that the randomisation succeeded), or that there were features in the baseline measures that needed to be explained or adjusted in the analyses.

6.5 Conclusions

In conclusion, this thesis has examined three types of misconduct: inappropriate sub-group analysis, selective reporting of outcomes, and data fabrication and falsification.

The experts who participated in the Delphi survey emphasised that inappropriate subgroups and selective reporting are very common and are the main source of distortion of conclusions from trials.

The evidence here shows that inadequate specification of sub-group analyses and unclear definition of the primary and secondary outcomes appears in supposedly high quality protocols. In the final reports of the trials in *the Lancet* protocols, which were peer reviewed for publication, inappropriate sub-group analysis and selective reporting did occur. Discrepancies between protocols and published reports were also common.

On the subject of fraud, a range of statistical techniques to detect fraud was listed and a selection of these was applied to a real dataset, and fraud was confirmed. Nevertheless, with suitable expertise, the use of computer techniques to fabricate data and with knowledge of the statistical techniques used for detection, a technically competent fraudster may be able to escape this route of detection. One might ask if these techniques to detect fraud could be routinely applied. However, I would recommend applying fraudulence tests sparingly, since routine use might lead to more sophisticated data simulation by fraudsters.

It is important to bear in mind that if these techniques indicate severe discrepancies, then the evidence for fraud is very strong. But, strictly speaking, will be conclusive only if other explanations cannot be ruled out.

6.6 Future Research

Following on from this thesis, comparison between pharmaceutical industry trials and academic trials, in terms of reporting sub-group analyses and selective reporting, would be interesting. For these trials, it would be useful to investigate the published reports, noting if a statistician were included as an author or member of the writing committee. It would be of interest to examine whether the input of a statistician improved the quality of the protocol, the analyses and the reporting.

The impact of a paper with flaws due to fraud or due to selective reporting can be assessed by how often its results are cited and used. In the case of the Singh trial, I looked to see how often it was quoted. One could take this further to see if the results influenced clinical guidelines when they were based on stated evidence.

A more complete exercise beyond the scope of this thesis, can be envisaged for selective reporting of outcomes. A suitably large sample of trial reports would be reviewed where results are based on sub-groups analyses or on reporting of more than one outcome. Reports where these are found would be assessed for the likelihood that the conclusion could be unreliable because of these aspects. The extent to which each report has been cited and to which it has been incorporated in systematic reviews would be evaluated. The importance given to it when cited and in systematic reviews, and how far this importance rests on selectively reported results, would be evaluated.

An alternative study might be to go back from the National Institute for Health and Clinical Excellence (NICE) guidelines and see how much the influence of specific studies on these guidelines could have been attributed to subgroups results.

A similar study to the first proposed above could be carried out for reports where fraud has been confirmed or is strongly suggested. However, fraud is concealed whereas selective reporting generally is not (or is less commonly concealed) and this means that a representative conclusion cannot be drawn.

It would also be of interest to examine further the extent to which fabricated trial data could be detected by statistical testing. To obtain a dataset that included realistic amounts of fabricated data, clinicians or statisticians could be provided with an existing dataset from a multi-centre trial and invited to fabricate the data for one or a few additional, imaginary centres. The statistical techniques outlined in Chapter 4 might then be applied

to this dataset to detect any suspicious data. This approach might provide further insights into the ability of statistical tests to detect fabricated data.

Reference:

1 Nimmo WS, editor. Joint Consensus Conference on Misconduct in Biomedical Research. *Proc R Coll Physicians Edinb* 2000; **30:**Supplement 7.

2 Martinson BC, Anderson MS, de Vries R. Scientists behaving badly. *Nature* 2005;435:737-738.

3 Bailar JC. Science, statistics, and deception. *Annals of Internal Medicine* 1986;104:259-60.

4 Singh RB, Dubnov G, Niaz MA, Ghosh S, Singh R, Rastogi SS, *et al.* Effect of an Indo-Mediterranean diet on progression of coronary artery disease in high risk patients (Indo-Mediterranean diet heart study): a randomised single-blind trial. *Lancet* 2002 360: 1455-1461.

5 Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004; 291:2457-65.

6 Godlee F. Publishing study protocols: making them more visible will improve registration, reporting and recruitment. *BMC News Views* 2001; 2:4.

7 Lassere M, Johnson K. The power of the protocol. *Lancet* 2002; 360:1620-1622.

8 Hawkey CJ. Journal should see original protocols for clinical trials. *BMJ* 2001; 323:1309.

9 Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomised trials funded by the Canadian Institutes of Health Research. *CMAJ* 2004; 17(7):735-740.

10 Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. *Medical Journal of Australia* 2004; 180(6):289-291.

11 Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005; 365(9454):176-186.

12 Smith R, Roberts I. patient safety requires a new way to publish clinical trials. *PLoS Clinical Trials* 2006; 1(1):e6 DOI: 10.1371/journal.pctr.0010006.

13 De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *Ann Intern Med.* 2004;141:477-8.

14 Chalmers I, Altman DG. How can medical journals help prevent poor medical research? Some opportunities presented by electronic publishing. *Lancet* 1999;353:490-493.

15 Gardner W, Lidz CW, Hartwig KC. Authors' reports about research integrity problems in clinical trials. *Contemporary clinical trials* 2005;26(2):244-51.

16 Geggie D.A survey of newly appointed consultants' attitudes towards research fraud. *Medical ethics* 2001;27(5):344-6.

17 Ranstam J, Buyse M, George SL, Evans S, Geller NL, Scherrer B, et al. Fraud in medical research: an international survey of biostatisticians. ISCB Subcommittee on Fraud. *Controlled clinical trials* 2000;21(5):415-27.

18 Adler M, Ziglio E, eds. Gazing into the oracle: The Delphi method and its application to social policy and public health. London, Kingsley, 1996.

19 Roberts-Davis M, Read SM. Clinical Role Clarification: using the Delphi method to establish similarities and differences between nurse practitioners and clinical nurse specialists *Journal of Clinical Nursing* 2001;10(1): 33-43.

20 Linstone H, Turoff M. The Delphi Method: Techniques and Applications. Massachusetts: Addison-Wesley, 1975.

21 Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing* 2000;32:1008-1015.

22 Crisp J, Pelletier D, Duffield C, Adams A, Nagy S. The Delphi Method?. *Nursing Research* 1997;46(2):116-118.

Appendix 1: Responses from round 2 Delphi survey

Please rate on a 5-point scale (1 = very UN-likely ..... 5 = very likely)

	STAGE 1: DESIGN	Likelihood to occur	Likelihood to distort the result
1	Biased literature review		
2	Fabricating references so that there looks to be a good case for a trial		
3	Fabricating clinical uncertainty so that there looks to be a good case for a trial		
4	Insufficient funds to conduct trial properly		
5	Lack of trial design experience		
6	Unnecessarily complex trial procedures		
7	Failure to specify in the protocol the main outcome measure		
8	Submissions to ethics committees that do not describe all trial procedures		
9	Failure to obtain ethics committee approval		
10	Non-independent Data Monitoring Committee		
11	Inappropriate methods for determining sample size		
12	Designing of studies with inadequate power		
13	Inappropriate or ambiguous inclusion and exclusion criteria		
14	Asking participants to agree to randomisation in the absence of equipoise		
15	Trial seeks to answer unimportant question		
16	Use of a cross-over where carry-over is expected		
17	Inappropriate use of 'equivalence' or 'non-inferiority designs'		
18	Failure to use random allocation		
19	Use of an unethical control group or intervention group		
20	Intentional use of non-optimum comparison treatment		
21	Use of placebo as comparison when other treatments exist		
22	Inadequate allocation concealment		
23	Lack of centralised follow-up		
24	Failure to employ evidence-based methods to minimise loss to follow-up		
25	Different follow-up schedules in arms		
26	Data collection forms designed to collect data that could never be analysed		
27	Inadequate blinding of outcome assessment		

28	Failure to anonymise case report form		
29	In an equivalence trial, choice of an inappropriate outcome measure		
30	Inappropriate timing of measurement of treatment effects		
31	Precision of measurement is avoided in an equivalence trial		
32	Centres chosen to optimise estimated treatment effects		
33	Failure to collect information on all clinically relevant end points		
34	Failure to follow good clinical practice		
35	Failure to use methods eg.(minimisation) to reduce chance imbalance		

Please rate on a 5-point scale (1 = very UN-likely ..... 5 = very likely)

	STAGE 2: CONDUCT	Likelihood to occur	Likelihood to distort the result
1	Lack of quality control mechanisms		
2	Inexperienced investigators		
3	Understaffing		
4	Insufficient support for training and supervision		
5	Improper consent procedures		
6	Failure to obtain informed consent (if required)		
7	Including patients who are unlikely to benefit when there are known harms		
8	Exerting too much pressure or giving excessive incentives for recruitment		
9	Fail to monitor adherence to inclusion and exclusion criteria		
10	Non inclusion of eligible patients particularly when randomisation is possible		
11	Non third-party randomisation		
12	Excessive data collection		
13	Allow (some or all) centres to depart from protocol		
14	Inconsistency of protocol adherence among centres		
15	Post-hoc changes in protocol		
16	Failure to have in place appropriate system for data monitoring		
17	Failure to use statistical methods to monitor data overall		
18	Selective withdrawals on basis of knowledge of allocation		
19	Treatment recognition in blinded trials		
20	Tampering with treatment packs so as to un-blind allocation		
21	Excessive financial rewards for following up patients		
22	Not achieving 100% follow-up		
23	Early termination of individual centre participation		
24	Failure to ensure security confidentiality of data		
25	Failure to apply advance data management systems		
26	Modifying clinical data to meet eligibility criteria		
27	Data falsification		
28	Data fabrication		
29	Failure to document and backup trial data		
30	Inadequate procedures for handling data (e.g. forms lost in post)		

Please rate on a 5-point scale (1 = very UN-likely ..... 5 = very likely)

	STAGE 3: ANALYSIS	Likelihood to occur	Likelihood to distort the result
1	Fail to specify a reasonable analysis plan in advance		
2	Ignore outliers		
3	Selective exclusion of "protocol violation outliers"		
4	Ignore data on side-effects		
5	Missing data ignored when informative		
6	Incorrectly imputing values for missing data		
7	Inexperienced statistician		
8	Fail to comply with a pre-specified analysis plan		
9	Use of inappropriate statistical methods		
10	Altering analysis methods until find significant result		
11	Lack of independent analysis		
12	Unplanned interim analysis		
13	Use of battery of methods of comparison to get the right answer		
14	Inappropriate sub-group analyses		
15	Sub-group analyses done without interaction tests		
16	Post-hoc analysis not admitted		
17	Deviation from intention to treat analysis		
18	Rely on biased comparisons as the primary analysis		
19	Use of primary outcome measure that was not pre-specified		
20	Misunderstanding / inadequate use of outcomes		
21	Having a large preference for a specific outcome		
22	Using a different primary endpoint from that specified in the protocol		
23	Excluding patients or results to exaggerate effects or remove adverse events		
24	Analysis completed by one person and not checked		
25	Inadequate model checking-eg proportional hazard assumption		
26	Failure to pay due regard to problems arising from multiplicity of various kinds		
27	Claiming equivalence by dint of failure to demonstrate a difference		
28	Most powerful analysis for efficacy; least powerful for safety		

29	Trial stopped for marketing and not scientific reasons		
30	Failure to account for 'clustering' issues (multi-level)		
31	Reducing data in a biased fashion		
32	Analysis conducted by the sponsor of the trial		
33	Selecting covariates to bias treatment effect in a particular direction		
34	Inappropriate analysis for example comparison of survival time by t-test		
35	Failure to account for all recorded results available in analysis		
36	Altering results in knowledge of allocation		

Please rate on a 5-point scale (1 = very UN-likely ..... 5 = very likely)

	STAGE 4: REPORTING	Likelihood to occur	Likelihood to distort the result
1	Not attributing authorship		
2	Gift authorship		
3	Selective reporting of outcomes in the abstract		
4	Failure to state in advance clinically relevant difference		
5	Not following CONSORT statement		
6	Inadequate description of methods (so that replication of study is impeded)		
7	Failure to describe changes to the intended analysis plan		
8	Claim an analysis is by "intention-to-treat" when it is not		
9	Reviewers: bias by reviewers because of personal interests		
10	Giving incomplete information about analyses with non significant results		
11	Failure to note multiple testing		
12	Insufficient time / resources available for in depth analysis		
13	Sub-group analyses may be indicated as pre-planed		
14	Unjustified extrapolation		
15	Over-interpretation of 'significant' findings in small trials		
16	Fixation with p –value rather than making use of confidence intervals		
17	Putting undue stress on results from sub-group analysis		
18	Selective reporting of positive results or omission of adverse events data		
19	Effect of (favoured treatment) reviewed in very favourable light		
20	Clinically important effect sizes may be declared to suit results		
21	Failure to report unfavourable results		
22	Poor use of figures which mislead / distort results		
23	Conclusion drawn that cannot be linked with evidence provided in report		
24	Problems encountered not reported		
25	Failure to report important facts		
26	Incomplete reporting		
27	Selective reporting based on p-values		
28	Selective reporting of (i) sub-groups (ii) outcomes (iii) time		

	points		
29	Report of sub-group without reference to wide study		
30	Pos hoc analyses reported as a main conclusion		
31	Report of single variable where multiple variables assessed and not reported		
32	No acknowledgment of sponsor		
33	Misleading citations in support of an argument		
34	Failure to acknowledge different results from other trials		
35	Multiple publications without reference to others		
36	Reporting under control of sponsor		
37	Negative or detrimental studies not published		
38	Failure to report results or long delay in reporting		
39	Competing interests not declared		
40	Manuscript not checked adequately by all co-authors		
41	Reporting of "events" before reporting of non events		
42	Redundant publication		

Appendix 2: Ethics Committee Approval

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

ETHICS COMMITTEE

APPROVAL FORM
Application number: **1095**

Name of Principal Investigator **Sanaa Al-Marzouki**

Department **Epidemiology and Population Health**

Head of Department **Pat Doyle**

Title: **Review of clinical trials protocols**

Approval of this study is granted by the Committee.

Chair T. W. Meade
Professor Tom Meade

Date 19.5.2004

Approval is dependent on local ethical approval having been received.

Any subsequent changes to the consent form must be re-submitted to the Committee.

Appendix 3: Published papers

Available online at www.sciencedirect.com

SCIENCE DIRECT

Contemporary Clinical Trials 26 (2005) 331–337

Contemporary Clinical Trials

www.elsevier.com/locate/conclintrial

The effect of scientific misconduct on the results of clinical trials: A Delphi survey

Sanaa Al-Marzouki*, Ian Roberts, Tom Marshall, Stephen Evans

Department of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, 49-51 Bedford Square, London, WC1B 3DP, United Kingdom

Received 27 August 2004; accepted 14 January 2005

Abstract

Objectives: To discover what types of scientific misconduct are most likely to influence the results of a clinical trial.
Design: Delphi survey of expert opinion with three rounds of consultation.
Setting: Non-industry clinical trial "community".
Participants: Experts identified from invitees to a previous MRC consultation on clinical trials. 32 out of the 40 experts approached agreed to participate.
Results: We identified thirteen forms of scientific misconduct for which there was majority agreement (>50%) that they would be likely or very likely to distort the results and majority agreement (>50%) that they would be likely or very likely to occur. Of these, the over-interpretation of 'significant' findings in small trials, selective reporting and inappropriate subgroup analyses were the main themes.
Conclusions: According to this expert group, the most important forms of scientific misconduct in clinical trials are selective reporting and the opportunistic use of the play of chance. Data fabrication and falsification were not rated highly because it was considered that these were unlikely to occur. Registration and publication of detailed clinical trial protocols could make an important contribution to preventing scientific misconduct.

Keywords: Scientific misconduct; Clinical trial; Delphi survey

* Corresponding author.
E-mail address: sanaa.al-marzouki@lshtm.ac.uk (S. Al-Marzouki).

1551-7144/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2005.01.011

1. Background

Scientific misconduct has been defined as behaviour by a researcher, whether intentional or not, that falls short of good ethical and scientific standards [1], and in particular can arise in the context of clinical trials. However, because the results from clinical trials are used to decide whether or not treatments are effective, decisions that may influence treatment choices for large numbers of patients, the prevention and detection of scientific misconduct in clinical trials is particularly important. Although any form of scientific misconduct can discredit the findings of a clinical trial, misconduct that distorts the estimate of the treatment effect or its precision is of special importance since it may lead to patients being given useless or harmful treatments or to patients being denied effective treatments. Nevertheless, there is currently little information about what types of scientific misconduct are most likely to distort the results of, or conclusions from, clinical trials.

This study used the Delphi methodology [2] among experts in clinical trials to provide an insight into what types of scientific misconduct are most likely to influence the results of a trial. The Delphi technique is a consensus method used to determine the extent of agreement on an issue. A panel of experts is asked to take part in a series of rounds to identify, clarify, refine, and finally to reach agreement on a particular issue. Because the panel do not meet, individuals can express their opinion without being influenced by others. In the Delphi method, anonymity of response enhances objectivity, the use of feedback through multiple iterations allows for a complete and thorough consideration and response, and the use of statistical analysis of the group response quantifies the strength of agreement and the pattern of agreement.

2. Methods

A group of 40 experts in clinical trials was assembled from the list of people invited to respond to the UK Medical Research Council (MRC) Clinical Trials for Tomorrow consultation [3]. Each expert was sent a letter explaining the aims and methods of the study and invited to take part in a Delphi survey with three rounds. Panel members were selected on the basis of their knowledge of the subject area and their willingness to be involved in research as is recommended when using the Delphi approach [4].

In the first round, each participating expert was asked to list, briefly and concisely, four suggestions about how scientific misconduct can arise in the design, conduct, analysis and reporting of a clinical trial. These suggestions were then collated and any duplicates were removed from the list in preparation for the second round.

In the second round, the list of collated suggestions was sent to each participant, whether or not they had responded to the first round. Participants were asked to rate each form of scientific misconduct on two dimensions: (1) the likelihood that it would occur in a clinical trial and (2) the likelihood that it would distort the results (i.e. have an effect on the magnitude of the treatment effect or its precision). Participants rated each suggestion on a five point scale from "very unlikely" to "very likely". A score of one indicated that the form of misconduct would be very unlikely to occur or would be very unlikely to distort the results. A score of five indicated that that form of misconduct would be very likely to occur or would be very likely to distort the results.

For round three, a list was prepared of all the forms of misconduct, showing the frequency distributions of the scores on both dimensions. Each participant's response in the second round was indicated under the appropriate number on the frequency distribution. Each participant was offered the

Table 1
Types of misconduct for which majority agreement was reached on the criterion of likely or very likely to distort the result, with percentages at this level of agreement and the percentage breakdown of respondents' views on the likelihood of occurrence

Types of misconduct	Percentage indicating likely or very likely to distort results	Likelihood to occur (%)				
		Very unlikely				Very likely
		1	2	3	4	5
Design						
Failure to use random allocation	92	12	68	16	0	4
Failure to specify in the protocol the main outcome measure	88	8	48	28	16	0
Inadequate allocation concealment	84	0	24	48	20	8
Different follow-up schedules in arms	80	8	40	52	0	0
Use of a cross-over where carry-over is expected	79	8	46	46	0	0
Intentional use of non-optimum comparison treatment	76	0	40	44	16	0
Precision of measurement is avoided in an equivalence trial	74	0	30	55	15	0
Inadequate blinding of outcome assessment	72	0	12	72	12	4
Inappropriate timing of measurement of treatment effects	60	4	20	68	8	0
In an equivalence trial, choice of an inappropriate outcome measure	56	0	28	56	16	0
Conduct						
Tampering with treatment packs so as to un-blind allocation	95	17	75	4	4	0
Selective withdrawals on basis of knowledge of allocation	92	8	52	28	12	0
Data falsification	92	64	32	4	0	0
Data fabrication	92	72	24	4	0	0
Treatment recognition in blinded trials	64	4	36	36	24	0
Post-hoc changes in protocol	52	0	20	56	20	4
Analysis						
Altering analysis methods until finding a significant result	100	4	28	60	8	0
Use of battery of methods of comparison to get the right answer	100	0	24	64	12	0
Altering results in knowledge of allocation	100	76	16	8	0	0
Excluding patients or results to exaggerate effects or remove adverse events	99	17	46	21	16	0
Use of primary outcome measure that was not pre-specified	96	12	48	28	12	0
Selecting covariates to bias treatment effect in a particular direction	96	16	40	32	12	0
Selective exclusion of "protocol violation outliers"	88	0	32	44	24	0

(continued on next page)

Table 1 (*continued*)

Types of misconduct	Percentage indicating likely or very likely to distort results	Likelihood to occur (%)				
		Very unlikely				Very likely
		1	2	3	4	5
Analysis						
Inappropriate subgroup analyses	88	0	8	28	48	16
Claiming equivalence by dint of failure to demonstrate a difference	88	0	8	42	38	12
Rely on biased comparisons as the primary analysis	87	0	57	30	13	0
Missing data ignored when informative	84	0	20	36	32	12
Using a different primary endpoint from that specified in the protocol	84	16	48	20	16	0
Post-hoc analysis not admitted	83	0	4	37	42	17
Trial stopped for marketing and not scientific reasons	83	0	32	45	14	9
Reducing data in a biased fashion	77	9	43	24	19	4
Incorrectly imputing values for missing data	76	4	36	44	12	4
Subgroup analyses done without interaction tests	75	0	0	25	50	25
Failure to account for 'clustering' issues (multi-level)	72	0	12	44	32	12
Fail to comply with a pre-specified analysis plan	68	0	32	48	16	4
Deviation from intention to treat analysis	68	0	8	60	24	8
Ignore data on side-effects	64	8	40	32	4	16
Fail to specify a reasonable analysis plan in advance	56	0	12	52	20	16
Use of inappropriate statistical methods	56	0	32	48	16	4
Analysis conducted by the sponsor of the trial	54	0	4	42	33	21
Inappropriate analysis for example comparison of survival time by *t*-test	52	4	32	56	8	0
Reporting						
Failure to report unfavourable results	100	0	8	56	20	16
Selective reporting of positive results or omission of adverse events data	96	0	8	32	24	36
Selective reporting based on *p*-values	92	0	0	20	64	16
Report of subgroup without reference to wide study	92	0	48	28	24	0
Pos hoc analyses reported as a main conclusion	92	0	32	44	24	0
Negative or detrimental studies not published	88	0	8	24	28	40
Over-interpretation of 'significant' findings in small trials	87	0	0	17	50	33
Putting undue stress on results from subgroup analysis	84	0	4	28	48	20
Selective reporting of (i) subgroups (ii) outcomes (iii) time points	80	0	4	32	40	24

Table 1 (*continued*)

Types of misconduct	Percentage indicating likely or very likely to distort results	Likelihood to occur (%)				
		Very unlikely				Very likely
		1	2	3	4	5
Reporting						
Report of single variable where multiple variables assessed and not reported	68	0	20	52	20	8
Failure to report results or long delay in reporting	68	0	16	24	24	36
Clinically important effect sizes may be declared to suit results	63	0	12	63	17	8
Poor use of figures which mislead/distort results	60	0	28	56	12	4
Unjustified extrapolation	58	0	17	46	33	4
Selective reporting of outcomes in the abstract	56	0	0	24	44	32
Conclusion drawn that cannot be linked with evidence provided in report	56	4	16	44	20	16
Reporting under control of sponsor	56	0	20	64	8	8
Claim an analysis is by "intention-to-treat" when it is not	52	4	24	48	12	12
Giving incomplete information about analyses with non significant results	52	0	4	40	32	24

opportunity to change his or her response in the light of the group's opinion by ticking a new value for the score or if they did not wish to change their opinion to tick the same number as before.

For the analyses, majority agreement was considered to have been achieved if more than half of the expert group gave the same score. Forms of misconduct for which there was majority agreement that it would be likely (score 4) or very likely (score 5) to distort the results of a clinical trial (these two scores being combined for this purpose) were listed with the distribution of opinions on the likelihood that this form of misconduct would actually occur.

3. Results

Of the 40 experts invited to take part, 32 agreed to participate in the study, of whom 26 (81%), 27 (84%), and 25 (78%) completed rounds one, two and three, respectively. The 26 respondents in round one generated a list of 84 suggestions for the design stage of clinical trials, 93 suggestions for the conduct stage, 88 suggestions for the analysis stage and 85 suggestions for the report stage. Editing and combining similar items reduced the list to 35 suggestions (design), 30 suggestions (conduct), 36 suggestions (analysis) and 42 suggestions (reporting).

At the end of the third round, there was majority agreement that 60 forms of scientific misconduct were likely or very likely to distort the results of a clinical trial (Table 1). The types of scientific misconduct for which there was majority agreement that they would be likely or very likely to distort the results and majority agreement that they would be likely or very likely to occur are shown in Table 2. Of

Table 2
Types of misconduct for which there was majority agreement (>50%) that they would be likely or very likely to distort the results, and that they would be likely or very likely to occur

Types of misconduct	Indicating likely or very likely to occur (%)
Over-interpretation of 'significant' findings in small trials	83
Selective reporting based on *p*-values	80
Selective reporting of outcomes in the abstract	76
Subgroup analyses done without interaction tests	75
Negative or detrimental studies not published	68
Putting undue stress on results from subgroup analysis	68
Inappropriate subgroup analyses	64
Selective reporting of (i) subgroups (ii) outcomes (iii) time points	64
Selective reporting of positive results or omission of adverse events data	60
Failure to report results or long delay in reporting	60
Post-hoc analysis not admitted	59
Giving incomplete information about analyses with non significant results	56
Analysis conducted by the sponsor of the trial	54

the 13 types of misconduct shown in Table 2 the most likely to occur was over-interpretation of 'significant' findings in small trials, while selective reporting and inappropriate subgroup analyses were the main themes, these being given as likely to occur by more than three quarters of the respondents.

4. Discussion

This study used an expert consensus approach to determine what experts in clinical trials believe are the most important forms of scientific misconduct in clinical trials. We had specified a-priori that the criterion for important in this context would be forms of misconduct believed to occur commonly and to distort the trial results. The results fall into two main categories: selective reporting of trial results and inappropriate subgroup analyses.

The main strength of the Delphi technique is that it optimises input from respondents and minimises the bias that can be encountered in face to face group interaction. In this case, each expert offered their opinions freely and without any peer pressure from others in the expert group. The expert panel was chosen because of their knowledge and experience in the conduct of clinical trials. There are no recommendations regarding the most appropriate panel size for the Delphi technique with typical panel sizes varying between 10 and several hundred members, nor are there any recommendations concerning the sampling techniques [5]. The Delphi technique is qualitative approach and although we believe it was an appropriate method for eliciting the opinions of the particular group of experts chosen, the extent to which our results can be generalised is open to question.

A limitation of this study was that some of the suggestions elicited in the first round were vague or ambiguous. As a result, it was difficult to accurately exclude duplicates and so the list that was used in the second and third Delphi rounds was somewhat repetitive. On the other hand, the consistent high ranking of selective reporting and inappropriate subgroup analyses does suggest that we have accurately identified the most important issues.

Although there has been considerable attention in the scientific literature on the problems of data fabrication and data falsification these were absent from our list of the most important forms of misconduct because there was majority agreement that these problems were very unlikely to occur. Our results suggest that selective reporting and the opportunistic use of the play of chance (inappropriate subgroup analyses) are more important considerations in ensuring that patients receive only effective treatments. Indeed, the two problems can be closely related. Multiple post-hoc subgroup analysis with selective reporting might easily result in authors making exaggerated subgroup claims about treatment effectiveness [6].

A publicly accessible inventory of trial protocols that include a clear description of the statistical analysis plan is a potential solution to the problems of selective reporting and subgroup analyses. Such an initiative is already underway and was given further impetus earlier this year when the UK NHS joined the worldwide effort to register clinical trials at inception [7]. This could be combined with rigorous and thorough statistical review in the peer review process of clinical trials to ensure that the subgroup analyses undertaken and reported were those specified in the protocol. Future research will need to assess the extent to which this initiative has been successful.

Acknowledgements

The authors thank the members of the expert panel for their support with this study: Kamran Abbasi, Douglas Altman, David Braunholtz, Phil Edwards, Diana Elbourne, Lelia Duley, Emma Hall, Julian Higgins, John Imeson, Chris Jennison, David Jones, Mike Kenward, Betty Kirkwood, John Lewis, Gordon Murray, Chris Palmer, Tim Peters Ruth Pickering, Peter Rothwell, Peter Sandercock, Stephen Senn, Haleema Shakur, Anne Truesdale, Paula Williamson.

References

[1] Nimmo WS. Joint consensus conference on misconduct in biomedical research. Proc R Coll Phys Edinb 2000; 30(Supplement 7).
[2] Linstone H, Turoff M. The Delphi Method: Techniques and Applications. Addison-Wesley; 1975.
[3] M.R.C. Clinical Trials for Tomorrow. London: Medical Research Council; 2003.
[4] Erlandson DA, Harris EL, Skipper BL, Allen SD. Doing Naturalistic Inquiry. A Guide to Methods. London: Whurr Publishers; 1993.
[5] Reid NG. The Delphi techniques: its contribution to the evaluation of professional practice. In: Ellis R, editor. Professional Competence and Quality Assurance in the Caring Professions. Beckenham, Kent: Croome-Helm, 1988.
[6] Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trials reporting: current practice and problems. Stat Med 2002;21:2917–30.
[7] Staessen JA, Bianchi G. Registration of trials and protocols. Lancet 2003;362:1009–10.

Are these data real? Statistical methods for the detection of data fabrication in clinical trials

Sanaa Al-Marzouki, Stephen Evans, Tom Marshall, Ian Roberts

Abstract

Objectives To test the application of statistical methods to detect data fabrication in a clinical trial.
Setting Data from two clinical trials: a trial of a dietary intervention for cardiovascular disease and a trial of a drug intervention for the same problem.
Outcome measures Baseline comparisons of means and variances of cardiovascular risk factors; digit preference overall and its pattern by group.
Results In the dietary intervention trial, variances for 16 of the 22 variables available at baseline were significally different, and 10 significant differences were seen in means for these variables. Some of these P values were extraordinarily small. Distributions of the final recorded digit were significantly different between the intervention and the control group at baseline for 14/22 variables in the dietary trial. In the drug trial, only five variables were available, and no significant differences between the groups for baseline values in means or variances or digit preference were seen.
Conclusions Several statistical features of the data from the dietary trial are so strongly suggestive of data fabrication that no other explanation is likely.

Introduction

Most statistical analyses of clinical trials are undertaken on the presumption that the data are genuine. Large accidental errors can be detected during data analysis,[1 2] but if people are trying to "make up" data they are likely to do it in such a way that it is not immediately obvious, avoiding any large discrepancies. Nevertheless, fraudulent data have particular statistical features that are not evident in data containing accidental errors, and several analytical methods have been developed to detect fraud in clinical trials.[3 4] The *BMJ* has taken a general interest in this field and has published a book on fraud and misconduct, now in its third edition, which has a chapter on statistical methods of detection of fraud.[5]

In this paper we use statistical techniques to examine data from two randomised controlled trials. In one trial, the possibility of scientific misconduct had been raised by *BMJ* referees, based on inconsistencies in calculated P values compared with the means, standard deviations, and sample sizes presented (see p 281). For comparison, we used the same methods to analyse a second trial for which there were no such concerns. We were not involved in either trial.

Methods

The trial about which doubts were raised (the diet trial) was a single blind, randomised controlled trial of the effects of a fruit and vegetable enriched diet in 831 patients with coronary heart disease, including patients with angina pectoris, myocardial infarction, or surrogate risk factors. Study participants were stated to be randomly allocated to the intervention diet (Group I, n=415) or to the control group, which was the patient's usual diet (Group C, n=416). The aim was to examine the effect of the intervention diet on risk factors for coronary artery disease after two years. We do not present data from the two year follow-up, because differences between groups could arise as a result of the interventions. After the reviewers had expressed suspicions about the integrity of the data, the *BMJ* requested the original trial data. These were provided by the trial's first author on handwritten sheets, which we entered on to computer, making appropriate checks to avoid transcription errors. The data are considered in the two randomised groups at baseline, Group I and Group C.

The second ("drug") trial was a randomised controlled trial of the effects of drug treatment in 21 750 patients with mild hypertension from 31 centres, from which we randomly selected five centres with 838 patients who had complete data for the selected variables. Study participants were randomly allocated to receive the drug (Group I, N=403) or a placebo (Group C, N=435). The aim was to determine whether drug treatment reduced the occurrence of stroke, death due to hypertension and coronary events in men and women aged 35-64 years, when followed for two years (again we do not present data from the follow-up). The drug trial data were provided by the trial investigators as computer files. The data are presented by treatment group (I or C) at baseline, using the same notation as for the diet trial. The variables in this study in common with the diet study are weight, diastolic blood pressure, systolic blood pressure, cholesterol measurements, and height. Further details of the methods and results from that trial have been published.[6]

Statistical methods

We conducted various tests on the baseline data of the randomised groups in both trials, looking for patterns that might indicate that the data in the diet trial were not generated by the normal process of making and recording individual measurements on a series of patients. We used the data from the drug trial for comparison, since we expected them to show patterns typical of data collected normally during a trial.

Using basic descriptive statistics and conventional statistical significance tests we compared the baseline data in the randomised groups in both trials. In a randomised trial, the data at baseline should be similar in the randomised groups. (The mean, the variability, the shape of the distribution of the data, and the pattern of data resulting from the methods of measurement must be similar since the groups can differ from one another only by chance factors.) This is the reason why in general, tests for statistical significance are not conducted at baseline in genuine trials. If such tests are carried out about one in 20 of such tests will be significant purely

See also p 281, and *Editorial* by Smith and Godlee

Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT
Sanaa Al-Marzouki *research student*
Stephen Evans *professor of pharmacoepidemiology, Medical Statistics Unit*
Tom Marshall *senior lecturer in medical statistics*
Ian Roberts *professor of epidemiology and public heath*
Correspondence to: S Evans stephen.evans@Lshtm.ac.uk

BMJ 2005;331:267-70

Table 1 Baseline variables in the two trials under comparison

	Diet		Drug	
	Intervention	Control	Intervention	Control
Weight (kg):				
Mean	65.74	65.59	70.27	70.08
Median	66	66	70	69
Mode	65	65	70	61
SD	7.89	7.64	11.6	12.4
Min	40	39	40	36
Max	87	85	111	120
Height (cm):				
Mean	165.1	165.28	162.1	162.6
Median	165	165	160	163
Mode	165	165	160	157
SD	6.91	3.93	9.22	9.14
Min	140	140	138	140
Max	179	178	190	188
Systolic blood pressure (mm Hg):				
Mean	134.2	131.9	184.4	184.6
Median	130	130	185	184
Mode	130	130	186	181
SD	18.5	16.9	12.2	12.9
Min	100	100	160	160
Max	200	195	209	210
Diastolic blood pressure (mm Hg):				
Mean	86.5	86.7	91.8	91.2
Median	86	85	92	91
Mode	80	85	101	90
SD	9.98	9.2	10.8	11.4
Min	60	60	46	50
Max	112	120	114	115
Cholesterol (mmol/l):				
Mean	5.46	5.43	6.68	6.57
Median	5.48	5.48	6.6	6.5
Mode	5.43	5.43	6.4	6.1
SD	0.352	0.296	1.26	1.21
Min	4.53	2.95	3.6	3.7
Max	6.52	6.00	12	10.8

by chance. We used *t* tests to compare the means of the randomised groups and F tests to compare the variances (standard deviations).

Data that are recorded (or invented) by people (as opposed to machines) tend to show preferences for certain numbers, such as rounding to the nearest 5 or 10. This is seen in the last recorded digit of numbers, and is called "digit preference." This digit preference should be similar between groups formed just by a chance process—randomisation. We used χ^2 tests to examine whether there was any tendency for the last digit to take on particular values and whether any observed digit preference was the same in the two groups created by randomisation. Digit preference can occur in all legitimate data based on human recording, but any pattern of this preference should be similar between groups formed using randomisation. We used SPSS, version 12.0.1 (Chicago, USA), for our data analysis.

Results

Table 1 shows descriptive summaries of variables common to both trials for both groups in each trial. The drug trial values show what might be expected in a randomised trial, but the diet trial shows notable differences in standard deviations for height and cholesterol measurements.

Table 2 shows for each trial the results of *t* and F tests, for differences in means and also in variances between the intervention and control groups at baseline for all available variables. In a genuine trial, correctly randomised, any such differences would be due to chance. Usually P values should not be quoted to greater precision than $P<0.001$, but because of the extreme nature of these P values, their exact value is given. In the diet trial, differences in variances were significant for 16 of the 22 variables that were available, as were 10 differences in means for these variables. Several of the P values were extraordinarily small. The expectation is that about 5% of such comparisons would have $P<0.05$, and extremely small P values should not occur. In the drug trial, none of the baseline means and none of the baseline variances showed statistically significant differences between the two groups, though only five variables were compared.

Table 3 shows the analysis of digit preference, assuming a uniform distribution of last digits. In the diet trial, all of the χ^2 values were highly significant, indicating that all the variables showed strong digit preference, although some preference is not unexpected. Digit preference was also evident for the results of a laboratory cholesterol test, which is unexpected since human estimation of the results is not usual. Measurements of height were not supplied for the diet trial (they were derivable from body mass index and weight for means, but this is not relevant for digit preference). In the drug trial, the χ^2 value was highly significant for height (indicating strong digit preference as might be expected) but not for any of the other measures Blood pressure measurement used a random zero machine, intended to remove digit preference. Table 4 shows the results of χ^2 testing for a difference in the pattern of digit preference between the two groups created by randomisation. This allows for the fact that digit preference can occur, but this should show a similar pattern in each of the randomised groups. In the diet trial, the final digit distributions are significantly different between the intervention group and the control group at baseline for all variables apart from cholesterol, fasting blood glucose, caffeine, carotene, and vitamin A. In the drug trial, the two randomised groups are far from being significantly different in terms of the final digit.

Discussion

The data from the diet trial have various anomalous statistical features that are not present in the data from the drug trial. These features are differences in means, and, even more noticeable, in variances at baseline and in differences in pattern of digit preference between randomised groups.

Magnitude of P values

These differences in the means and variances between baseline variables in the diet trial indicate that the two groups simply cannot have been formed as a result of random allocation as the authors claim. The magnitude of the P values derived from *t* tests of these differences for several variables is not compatible with a chance effect. One or two variables might show a small effect, but several of these P values are extreme.

Table 2 Baseline comparison of the two intervention groups, diet trial and drug trial

	Diet trial				Drug trial			
	Levene's F test for equality of variances		t test for equality of means		Levene's F test for equality of variances		t test for equality of means	
	F	Significance	t	Significance (two tailed)	F	Significance	t	Significance (two tailed)
Height	71.15	1.4×10^{-16}	−0.508	0.612	0.054	0.82	0.82	0.411
Weight	0.204	0.652	0.284	0.776	2.46	0.12	−0.227	0.82
Systolic blood pressure	4.81	0.029	1.89	0.06	2.45	0.12	0.206	0.84
Diastolic blood pressure	4.366	0.037	−0.27	0.788	0.89	0.35	−0.679	0.497
Cholesterol	28.77	1×10^{-7}	1.19	0.235	0.27	0.61	−1.22	0.22
Fasting blood glucose	8.21	0.004	−0.57	0.566	—	—	—	—
Total cholesterol	0.043	0.835	−0.35	0.729	—	—	—	—
Triglycerides	21.98	3×10^{-6}	0.484	0.628	—	—	—	—
Energy	0.98	0.322	−1.57	0.118	—	—	—	—
Total carbohydrate	1.97	0.161	0.236	0.814	—	—	—	—
Complex carbohydrate	12.86	0.0004	14.8	6×10^{-44}	—	—	—	—
Protein	15.18	0.0002	5.02	6×10^{-7}	—	—	—	—
Fat	20.5	7×10^{-6}	−2.88	0.004	—	—	—	—
Saturated	15.2	0.0001	3.9	0.0002	—	—	—	—
Fibre	94.23	4×10^{-21}	−8.47	2×10^{-16}	—	—	—	—
Soluble fibre	10.13	0.002	−6.95	7×10^{-12}	—	—	—	—
Caffeine	2.41	0.121	0.957	0.339	—	—	—	—
Salt	39.72	5×10^{-10}	−.377	0.706	—	—	—	—
Vitamin C	0.007	0.931	−5.6	3×10^{-8}	—	—	—	—
Carotene	51.06	2×10^{-12}	29.8	2×10^{-133}	—	—	—	—
Vitamin E	25.7	5×10^{-7}	5.9	5×10^{-9}	—	—	—	—
Vitamin A	51.42	2×10^{-12}	4.49	8×10^{-6}	—	—	—	—

Similarly, the significant difference in the pattern of digit preference between the randomised groups provides additional evidence that this is not a truly randomised trial.

Randomisation process

If this is not a randomised trial then how did these data arise? One possibility is that the data themselves are genuine but that the randomisation process has been subverted. This might explain, for example, some of the differences between the means of the variables at baseline. Had there been subversion of the randomisation process, in order for example to create differences between the groups at baseline, then smaller differences would have occurred and would also have been more consistent between the variables that are medically related—such as the different meas-

Table 3 χ^2 value (with P value) for the final digit at baseline, diet trial and drug trial

	Diet trial*		Drug trial	
	Intervention	Control	Intervention	Control
Height	—	—	239 (1.8×10^{-46})	251 (7.2×10^{-49})
Weight	128 (4×10^{-23})	23 (0.00655)	7.3 (0.60)	6.5 (0.69)
Systolic blood pressure	1796 (U)	1470 (U)	7.6 (0.58)	9.1 (0.43)
Diastolic blood pressure	763 (2×10^{-158})	820 (1×10^{-170})	8.1 (0.52)	13.8 (0.13)
Cholesterol	554 (2×10^{-113})	430 (6×10^{-87})	16.23 (0.062)	5.76 (0.76)
Fasting blood glucose	478 (4×10^{-97})	538 (5×10^{-110})	—	—
Total cholesterol	1053 (6×10^{-221})	1522 (U)	—	—
Triglycerides	642 (2×10^{-132})	963 (2×10^{-201})	—	—
Energy	2151 (U)	2630 (U)	—	—
Total carbohydrates	207 (1×10^{-39})	927 (7×10^{-194})	—	—
Complex carbohydrates	231 (1×10^{-44})	939 (3×10^{-196})	—	—
Protein	54 (2×10^{-8})	251 (5×10^{-49})	—	—
Fat	229 (2×10^{-44})	437 (2×10^{-88})	—	—
Saturated	123 (4×10^{-22})	98 (4×10^{-17})	—	—
Fibre	263 (2×10^{-51})	1127 (9×10^{-237})	—	—
Soluble fibre	273 (1×10^{-53})	1086 (6×10^{-228})	—	—
Caffeine	613 (3×10^{-126})	694 (1×10^{-143})	—	—
Salt	288 (9×10^{-57})	301 (2×10^{-59})	—	—
Vitamin C	304 (5×10^{-60})	411 (6×10^{-83})	—	—
Carotene	1470 (U)	1156 (5×10^{-243})	—	—
Vitamin E	118 (3×10^{-21})	101 (8×10^{-18})	—	—
Vitamin A	705 (6×10^{-146})	799 (3×10^{-166})	—	—

The χ^2 value has 9 degrees of freedom.
* U means that the P value is too small for calculation.

Table 4 χ^2 value (with P value) for the final digit at the baseline in the diet and drug trials between the two randomised groups

	Diet trial		Drug trial	
	χ^2test (P value)	df	χ^2 test (P value)	df
Height	—	—	5 (0.83)	9
Weight	36 (3×10^{-5})	9	10 (0.31)	9
Systolic blood pressure	26 (0.00019)	6	7 (0.69)	9
Diastolic blood pressure	16 (0.046)	8	10 (0.38)	9
Cholesterol	13 (0.182)	9	7 (0.60)	9
Fasting blood glucose	12 (0.2)	9	—	
Total cholesterol	46 (5×10^{-7})	9	—	
Triglycerides	48 (3×10^{-7})	9	—	
Energy	16 (0.064)	9	—	
Total carbohydrate	154 (2×10^{-28})	9	—	
Complex carbohydrate	135 (1.4×10^{-24})	9	—	
Protein	43 (2×10^{-6})	9	—	
Fat	40 (6.4×10^{-6})	9	—	
Saturated	15 (0.08)	9	—	
Fibre	157 (8×10^{30})	8	—	
Soluble fibre	175 (6.5×10^{33})	9	—	
Caffeine	15 (0.059)	8	—	
Salt	28.5 (0.001)	9	—	
Vitamin C	18 (0.03)	9	—	
Carotene	10 (0.266)	8	—	
Vitamin E	20 (0.017)	9	—	
Vitamin A	9.5 (0.4)	9	—	

The degrees of freedom are less than 9 when one or more digits do not appear.

ures of cholesterol that show entirely different patterns between the groups. As it is, some are extreme and others are no different between the groups. What is more difficult to explain on the basis of subversion of the randomisation is the difference in the variability at baseline. Here we have highly significant differences in some variables both for the variances and the means, whereas for height, complex cholesterol, and triglyceride, there are highly significant differences in the variances but not in the means. Had there been a tendency to put patients with, say, higher blood pressures into one group, then we might have found significant differences in the mean values but with no difference in variance. However, we did not find this. Furthermore, no clear differences were apparent in the means for variables that would be readily available to a physician or health professional at the time of recruitment.

What is already known on this topic

Data fabrication is a rare form of scientific misconduct in clinical trials, but when it does occur it has serious consequences

Most papers are published without their data being independently verified, and there have been calls for data to be made available for scrutiny

Statistical methods for the detection of misconduct have been described, but few examples of their application have been published

It has been stated that statistical methods alone cannot prove data fabrication

What this study adds

Statistical methods can be applied to detect large scale fabrication of data in a randomised trial where data are available

Certain patterns of data are incompatible with randomisation, especially when a trial is "blind"

This paper shows the fabrication or falsification of data in a particular trial

Digit preference

Digit preference in itself is not evidence of misconduct. It is conceivable that the different patterns of digit preference between the two randomised groups may have arisen had one person recorded data for the treatment group and another recorded data for the control group. However, it is claimed that the trial was single blind, meaning that those recording data should not know to which group patients had been allocated. We would not expect differences therefore in digit preference between the randomised groups. But perhaps the trial was not single blind as described, and those recording the data were separated into groups according to whether they were dealing with patients allocated to either treatment or control. This could lead to differences in digit preference between randomised groups for variables where a human element of judgment was required. This would still not explain the differences in means and variances between the two groups since the effect of digit preference on the means and variances would only be slight. The combination of the differences in means, variances, and digit preference between the randomised groups is strong evidence that data fabrication took place in the diet trial.

Conclusion

We conclude that the data from the diet trial were either fabricated or falsified and that the strength of the evidence is such that appropriate steps should be taken to deal with this matter.

We thank Tom Meade who, on behalf of the Medical Research Council, provided the data for the drug trial and Richard Smith for his encouragement to examine further the data from the diet trial. The *BMJ* provided the data from the diet trial, which were supplied by the original author for further investigation of these data.

Contributors: SE and SAM had the ideas for the analysis, and SAM, SE, TM, and IR all contributed to the planning, conduct, and writing of the paper. SAM planned and carried out the statistical analyses. SAM and SE are jointly responsible for the overall content as guarantors. There are no other contributors.

Competing interests: None declared.

Funding: None.

1 Armitage P, Berry G. *Statistical methods in medical research.* 3rd ed. Oxford: Blackwell Scientific, 1994:386-401.
2 Altman DG. *Practical statistics for medical research.* London : Chapman and Hall, 1991:122-143.
3 Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B, et al. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Stat Med* 1999;18:3435-51.
4 Taylor RN, McEntegart DJ, Stillman EC. Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. *Drug Inform J* 2002;36:115-25.
5 Evans S. Statistical aspects of the detection of fraud. In: Lock S, Wells F, Farthing M, eds. *Fraud and misconduct in medical research.* 3rd ed. London: BMJ Publishing Group, 2001:186-204.
6 Medical Research Council Working Party. MRC trial of treatment of mild hypertension: principal result. *BMJ* 1985;291:97-104.

(Accepted 15 July 2005)